Advanced Spatio-Temporal Neural Networks for Malaria Prevalence Forecasting in Sub-Saharan Africa

Vikalp Srivastava[1], Manoj Kumar[1,*], Sanika Vaidya[2]
[1]School of Engineering and Science
Indian Institute of Technology Madras, Zanzibar
Tanzania
[2] Department of Mathematics
Institute of Chemical Technology, Mumbai
∗Corresponding author's email: manoj@iitmz.ac.in

Abstract

Accurate malaria prevalence forecasting is critical for timely interventions in Sub-Saharan Africa, where the disease remains a major health burden. Traditional regression models often struggle to capture the complex interactions between the epidemiological, environmental, and spatial factors that drive malaria transmission. This study develops and compares two distinct deep learning approaches for predicting the *Plasmodium falciparum* parasite rate in children (PfPR_{2-10}), using historical prevalence data (1900–2015) and climate variables—temperature and precipitation—sourced from NASA's POWER project. The first framework is a Spatiotemporal Transformer developed to capture intricate, local dependencies between covariates. This is contrasted with a second approach based on the Patch-based Time Series Transformer (PatchTST), which includes a standard implementation for point forecasting. Both models were further extended with uncertainty estimation using Monte Carlo Dropout, providing probabilistic forecasts in addition to point predictions. Our findings highlight a key architectural trade-off: for this dataset, characterized by short time-series and strong cross-feature interactions, the Spatiotemporal Transformer yielded higher predictive accuracy. Conversely, while demonstrating lower point-forecast accuracy in this context, both models successfully provided valuable uncertainty estimates for their predictions. This comparative analysis provides a nuanced perspective, offering decision-makers a choice between a highly accurate model for direct forecasting and probabilistic models for risk assessment and strategic planning.

1 Introduction

Malaria remains one of the most persistent and devastating public health challenges in Sub-Saharan Africa, affecting millions each year and placing a substantial burden on healthcare systems, economies, and communities. According to the World Health Organization's latest reports [9], the region accounts for approximately 95% of global malaria cases and 96% of malaria deaths, with children under five years of age being disproportionately affected. A key aspect of effective malaria control lies in the ability to accurately forecast disease prevalence across different temporal and spatial scales [6]. Such forecasting capabilities enable policymakers, health organizations, and international aid agencies to allocate resources more efficiently, deploy preventative measures proactively, and respond to potential outbreaks before they reach critical

levels. However, malaria transmission is influenced by a complex and dynamic interplay of epidemiological, environmental [12], and spatial factors, making accurate prediction a challenging and multifaceted task [5].

Environmental factors such as temperature and precipitation significantly influence mosquito breeding cycles and survival rates [7]. Spatial factors including population density, urbanization levels, access to healthcare, and economic development impact both disease transmission patterns and treatment outcomes. Additionally, biological factors such as vector species distribution, parasite drug resistance patterns, and host immunity levels add further complexity to the prediction challenge.

Traditional regression models often fall short in capturing these complex interactions between the epidemiological, environmental, and spatial factors that drive malaria transmission due to their reliance on linear assumptions and limited capacity to model complex interactions between variables [1–3]. Early machine learning approaches, while more flexible, typically treat spatial and temporal dimensions as independent features, failing to capture the inherent spatio-temporal correlations that characterize vector-borne diseases like malaria.

Recent advances in machine learning and deep learning have shown significant promise in infectious disease modeling and diagnosis. [11] explored interpretable ML frameworks for malaria prediction, addressing the crucial need for transparency in health applications. Complementing this, [8] developed an ensemble AI-enabled MIoT system for automated malaria parasite detection, while [10] introduced MalNet-DAF, a dual-attentive fusion deep model that enhances classification performance. Similarly, [4] demonstrated the utility of YOLO-mp for real-time malaria parasite detection and counting, advancing automated diagnostics. Beyond malaria, [14] applied deep epidemiological modeling with black-box knowledge distillation for accurate COVID-19 predictions, highlighting the broader applicability of AI in epidemiology.

This study develops and compares two distinct deep learning approaches for predicting the *Plasmodium falciparum* parasite rate among children aged 2 to 10 (PfPR_{2-10}), using historical prevalence data (1900–2015) and climate variables—temperature and precipitation—sourced from NASA's POWER project. The first framework is a Spatiotemporal Transformer developed to capture intricate, local dependencies between covariates. This is contrasted with a second approach based on the Patch-based Time Series Transformer (PatchTST), which includes a standard implementation for point forecasting and a Monte Carlo Dropout variant for quantifying uncertainty.

2 Methodology

2.1 Problem Formulation and Mathematical Framework

Our malaria forecasting problem can be formally defined as a supervised spatio-temporal regression task where we aim to predict future malaria prevalence given historical multi-modal data. Let $\mathcal{D} = \{(\mathbf{X}_i^{(T)}, \mathbf{X}_i^{(S)}, y_i)\}_{i=1}^N$ represent our dataset of N samples, where:

$$\begin{aligned}
\mathbf{X}^{(T)} &\in \mathbb{R}^{L \times d_T} \quad \text{(temporal feature sequence)} \\
\mathbf{X}^{(S)} &\in \mathbb{R}^{d_S} \quad \text{(spatial feature vector)} \\
y &\in [0, 1] \quad \text{(normalized } \text{PfPR}_{2-10} \text{ prevalence rate)},
\end{aligned}$$

where L represents the sequence length, d_T denotes the total number of time-varying features (including epidemiological indicators, derived patterns, and climate variables), and d_S represents the static spatial dimensions (geographic coordinates and location-specific statistics).

Our objective is to learn a function $f : \mathbb{R}^{L \times d_T} \times \mathbb{R}^{d_S} \to [0, 1]$ such that:

$$\hat{y} \approx f(\mathbf{X}^{(T)}, \mathbf{X}^{(S)}) \tag{2.1}$$

The learning objective minimizes the expected prediction error:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{X}^{(T)}, \mathbf{X}^{(S)}, y) \sim \mathcal{D}} \left[\ell(f(\mathbf{X}^{(T)}, \mathbf{X}^{(S)}; \theta), y) \right] \tag{2.2}$$

where $\ell(\cdot, \cdot)$ is the loss function, θ represents the learnable parameters, and the expectation is taken over the data distribution.

2.2 Dataset Description and Characteristics

We develop a spatio-temporal malaria prediction framework using two main datasets:

Malaria Prevalence Data: From the Malaria SSR Database [13], containing 28,632 geo-referenced measurements across Sub-Saharan Africa. Temporal features include year, month, age range (LoAge, UpAge), sample size (Ex, Pf), seasonal/trend indicators, and lagged variables. Spatial features include coordinates, administrative boundaries, survey methods, and derived spatial interactions.

Climate Data: Monthly meteorological data from NASA's POWER database (https://power.larc.nasa.gov) at $0.5^\circ \times 0.625^\circ$ resolution, including temperature (T2M), precipitation (PRECTOTCORR), climate interactions (products, ratios, non-linear transforms), and statistical summaries. Data are spatially and temporally aligned with prevalence records via nearest-neighbor interpolation. The target, PfPR_{2-10}, is the age-standardized prevalence of *P. falciparum* in children aged 2–10 years, normalized to $[0, 1]$.

2.3 Data Pre-processing and Feature Engineering

Our pipeline addresses missing data, feature enrichment, and sequence preparation.

Missing Values: Numerical features were median-imputed, categorical features mode-imputed.

Feature Engineering: Temporal (cyclical month encoding, year normalization, lags), spatial (lat-lon interactions, equator distance, hemisphere), climate (temp–precip products, ratios, non-linear terms), and statistical (3-period moving averages, location-specific means/SDs) features were derived.

Sequence & Grouping: Sliding windows of $L = 6$ observations captured seasonal patterns. Spatial grouping by country and admin codes preserved geographic relationships.

Normalization: Numerical features were standardized (zero mean, unit variance); target scaled to $[0, 1]$.

3 Incorporating Climate Data

We formulate malaria prevalence prediction as a supervised spatiotemporal forecasting problem. The model is provided a sequence of time-varying features $\mathbf{X}^{(T)}$ and a corresponding static spatial feature vector $\mathbf{X}^{(S)}$. The objective is to predict the malaria prevalence y.

The temporal feature sequence $\mathbf{X}^{(T)} \in \mathbb{R}^{L \times d_T}$ contains all time-varying inputs, including epidemiological features (e.g., month, age, survey data), derived patterns (e.g., lags, moving averages), and raw climate variables (e.g., monthly temperature and precipitation). The static spatial context is represented by $\mathbf{X}^{(S)} \in \mathbb{R}^{d_S}$, which encompasses geographic coordinates and location-specific statistics, such as the long-term mean and standard deviation of the climate variables for a particular location.

The prediction task aims to learn a mapping function $f : \mathbb{R}^{L \times d_T} \times \mathbb{R}^{d_S} \to [0, 1]$ such that:

$$\hat{y} \approx f(\mathbf{X}^{(T)}, \mathbf{X}^{(S)}) \tag{3.3}$$

This formulation allows the model to leverage both short-term temporal patterns and long-term spatial and environmental contexts for improved forecasting accuracy.

3.1 Enhanced Spatio-Temporal Transformer (ESTT) Architecture

Our proposed architecture extends standard transformer encoders with specialized components designed for spatiotemporal climate-disease modeling. The model consists of four primary modules: temporal encoding, transformer processing, spatial integration, and fusion-based prediction.

Temporal Feature Projection: The input temporal sequence $\mathbf{X}^{(T)} \in \mathbb{R}^{L \times d_T}$ (where $L = 6$ for sequence length) undergoes linear projection to the model's hidden dimension, followed by layer normalization and nonlinear activation:

$$\mathbf{X}_{\text{proj}} = \text{Dropout}(\text{GELU}(\text{LayerNorm}(\mathbf{X}^{(T)}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}})))$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_T \times d_{\text{model}}}$ and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{d_{\text{model}}}$ are learnable parameters.

Positional Encoding: To incorporate temporal ordering information, we add learnable positional embeddings to the projected features:

$$\mathbf{H}_0 = \mathbf{X}_{\text{proj}} + \mathbf{P}_{1:L} \tag{3.4}$$

where $\mathbf{P} \in \mathbb{R}^{N_{pos} \times d_{\text{model}}}$ represents the full matrix of position embeddings ($N_{pos} = 100$) and $\mathbf{P}_{1:L}$ denotes the first L embeddings corresponding to the sequence length.

Multi-Layer Transformer Encoding: The temporally encoded features pass through 3 transformer encoder layers with a **norm-first** architecture and residual connections:

$$\mathbf{H}_\ell = \text{TransformerEncoderLayer}_\ell(\mathbf{H}_{\ell-1}), \quad \ell = 1, \ldots, 3 \tag{3.5}$$

Each transformer layer incorporates multi-head self-attention ($n_{\text{heads}} = 8$) with GELU activations and increased feedforward dimensions ($4 \times d_{\text{model}}$) to capture complex temporal dependencies.

Enhanced Spatial Feature Processing: Spatial features $\mathbf{X}^{(S)} \in \mathbb{R}^{d_S}$ (including geographic coordinates, location statistics, and derived spatial interactions) undergo projection and multi-layer processing:

$$\mathbf{S}_{\text{proj}} = \text{Linear}(\mathbf{X}^{(S)}) \tag{3.6}$$

$$\mathbf{H}_{spatial} = \text{SpatialMLP}(\text{LayerNorm}(\mathbf{S}_{\text{proj}})) \tag{3.7}$$

The spatial MLP consists of two fully connected layers with layer normalization, GELU activation, and dropout, reducing the feature dimensionality to $d_{\text{model}}/2$.

Attention-Based Temporal-Spatial Fusion: We employ a multi-head self-attention layer ($n_{\text{heads}} = 4$) to refine the temporal representations before pooling:

$$\mathbf{H}_{\text{fused}} = \text{MultiHeadAttention}(\mathbf{H}_3, \mathbf{H}_3, \mathbf{H}_3) \tag{3.8}$$

$$\mathbf{h}_{temp} = \frac{1}{L} \sum_{t=1}^{L} \mathbf{H}_{\text{fused}}[t, :] \tag{3.9}$$

Enhanced Prediction Module: The final prediction is generated by a deep MLP that takes the concatenated temporal and spatial feature vectors:

$$\hat{y} = \text{PredictionMLP}([\mathbf{h}_{temp}; \mathbf{H}_{spatial}]) \tag{3.10}$$

The prediction MLP incorporates four layers with dimensions $(d_{\text{model}} + d_{\text{model}}/2) \to d_{\text{model}} \to d_{\text{model}}/2 \to d_{\text{model}}/4 \to 1$, utilizing layer normalization, GELU activation, and dropout for robust training.

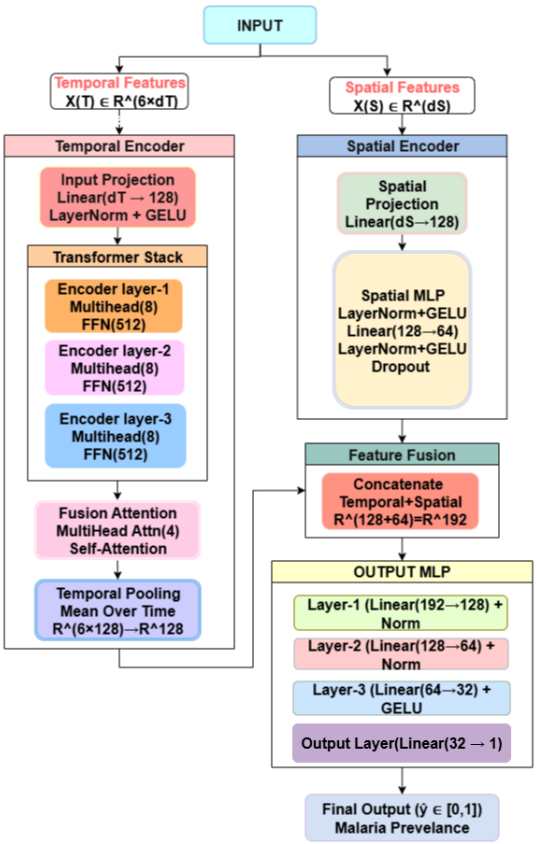


Figure 1: Enhanced Spatiotemporal Transformer Architecture for Climate-Disease Modeling

4 PatchTST-Based Spatiotemporal Forecasting Architecture

We implement an enhanced PatchTST architecture specifically adapted for spatiotemporal malaria forecasting, leveraging patch-based decomposition to efficiently process heterogeneous epidemiological and climatic time series data.

The key innovation of our approach lies in decomposing time series into patches and processing each feature channel independently, enabling specialized learning for different data modalities (epidemiological vs. climatic variables) while maintaining computational efficiency.

4.1 Patch-Based Decomposition

Given the multivariate time series $\mathbf{X}^{(T)} \in \mathbb{R}^{L \times d_T}$, the model first separates it into d_T channels. For each channel $\mathbf{X}^{(T,c)} \in \mathbb{R}^L$, patches of length P and stride S are extracted:

$$\mathbf{Patch}_j^{(c)} = \mathbf{X}^{(T,c)}[jS : jS + P], \quad j = 0, \ldots, \lfloor \frac{L-P}{S} \rfloor$$

Each patch is flattened, linearly projected to $\mathbb{R}^{d_{\text{model}}}$, and added with learnable positional encodings.

4.2 Channel-Independent Processing

Each feature channel $\mathbf{X}^{(T,c)} \in \mathbb{R}^L$ is patch-embedded and passed through a dedicated Transformer encoder with multi-head self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Outputs from all d_T channels are concatenated and linearly projected to form the aggregated temporal representation $\mathbf{h}_{\text{agg}}$.

4.3 Spatial Feature Integration

Spatial attributes $\mathbf{X}^{(S)} \in \mathbb{R}^{d_S}$ (e.g., coordinates, climate statistics) are projected via:

$$\begin{aligned} \mathbf{S}_{\text{proj}} &= \text{Linear}(\mathbf{X}^{(S)}) \\ \mathbf{H}_{spatial} &= \text{SpatialMLP}(\text{LayerNorm}(\mathbf{S}_{\text{proj}})) \end{aligned} \tag{4.11}$$

The SpatialMLP uses GELU activation and dropout.

4.4 Spatiotemporal Fusion and Prediction

The aggregated temporal representation $\mathbf{h}_{\text{agg}}$ is first refined using a self-attention mechanism to capture the most salient temporal patterns. The resulting features, $\mathbf{h}_{temp}$, are then fused with the processed spatial features $\mathbf{H}_{spatial}$ via concatenation.

1. **Temporal Feature Refinement (Self-Attention):**

$$\mathbf{h}_{temp} = \text{SelfAttention}(\mathbf{h}_{\text{agg}})$$

2. **Spatiotemporal Fusion (Concatenation) and Final Prediction:**

$$\hat{y} = \text{PredictionMLP}([\mathbf{h}_{temp}; \mathbf{H}_{spatial}])$$

4.5 Training

Smooth L1 loss with L2 regularization:

$$\mathcal{L}_{total} = \mathcal{L}(y, \hat{y}) + \lambda \sum_{\theta} \theta^2$$

Optimization uses AdamW with cosine annealing warm restarts and gradient clipping for stability.

This novel adaptation of PatchTST for spatiotemporal malaria forecasting represents a significant advancement in epidemiological modeling, combining the efficiency of patch-based processing with the sophistication of attention mechanisms specifically tailored for climate-enhanced disease prediction in resource-constrained settings

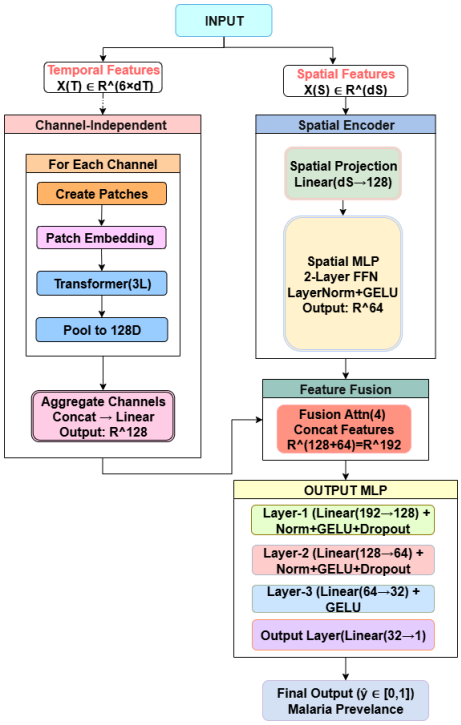


Figure 2: PatchTST Architecture for Climate-Disease Modeling

5 Experimental Results

In this section, we have done comparative analysis of three different statistical regression models, Poisson, Binomial, Linear Regression (comaprision is shown in Fig. 3), and advanced spatiotemporal models spatiotemporal transformer, PatchTST to predict the number of malaria parasites detected ('Pf') based on geographical coordinates ('Lat' and 'Long'). Table 1 presents a comprehensive performance comparison between different models.

Methodology	RMSE	MAE	R^2
Spatiotemporal Transformer	**9.688**	**6.497**	**0.841**
PatchTST	**17.448**	**12.535**	**0.486**
Poisson Regression	124.4111	41.6535	0.0065
Binomial Regression	146.2883	38.7158	-0.3736
Linear Regression	124.3484	41.6155	0.0075

Table 1: Comparision of traditional and proposed models.

5.1 Multi-Country Evaluation

We also evaluate our spatiotemporal transformer architecture across five sub-Saharan African countries with diverse malaria transmission patterns and geographic characteristics: Congo, Tanzania, Nigeria, Mozambique, and Uganda. Each country presents unique epidemiological challenges, ranging from high-burden equatorial regions to seasonal transmission zones, enabling comprehensive assessment of model generalizability.

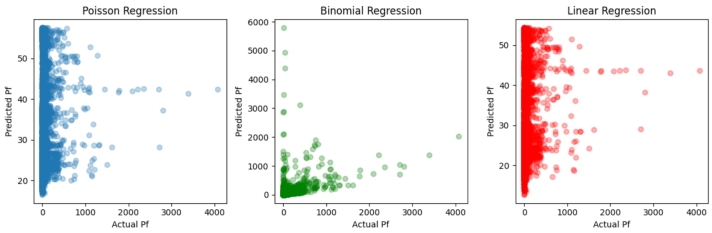


Figure 3: Actual vs. Predicted values for the Poisson, Binomial, and Linear Regression models.

5.2 Performance Analysis

The results demonstrate consistent improvements of the climate-enhanced model across multiple performance metrics and geographic regions, validating the effectiveness of our spatiotemporal transformer architecture for malaria prediction.

Overall Performance Gains: Across all five countries, the climate-enhanced model achieves an average MAE reduction of 3.6% (from 8.718 to 8.404), indicating improved prediction accuracy. The average R^2 improvement of 1.35% (from 0.742 to 0.752) demonstrates slightly enhanced explanatory power, while an RMSE reduction of 1.8% suggests modest improvements in handling prediction variance.

2*Country	Climate Model			Baseline Model (Binomial)		
	MAE	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**
Congo	30.205	-1.218	40.639	30.198	-1.217	40.629
Tanzania	14.049	-0.475	24.406	14.043	-0.471	24.372
Nigeria	30.634	-1.732	38.393	30.688	-1.740	38.449
Mozambique	39.652	-2.652	46.454	39.737	-2.666	46.545
Uganda	33.887	-1.979	41.512	33.956	-1.988	41.577
Average	**29.285**	**-1.611**	**38.281**	**29.324**	**-1.616**	**38.314**

Table 2: Performance Comparison: Climate Model vs Baseline Model (Binomial) across five countries

2*Country	Climate Model			Baseline Model (ESTT)		
	MAE	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**
Congo	5.907	0.818	8.976	7.590	0.802	9.382
Tanzania	7.978	0.771	11.063	7.808	0.780	10.841
Nigeria	10.045	0.718	13.001	10.240	0.687	13.692
Mozambique	9.455	0.757	12.378	9.403	0.755	12.431
Uganda	8.633	0.694	11.344	8.549	0.688	11.456
Average	**8.404**	**0.752**	**11.352**	**8.718**	**0.742**	**11.560**

Table 3: Performance Comparison: Climate Model vs Baseline Model (ESTT) across five countries

2*Country	Climate Model			Baseline Model (PatchTST)		
	MAE	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**
Congo	16.221	0.063	20.389	16.096	0.068	20.333
Tanzania	14.984	0.310	19.210	14.962	0.305	19.287
Nigeria	17.289	0.262	21.019	17.446	0.257	21.097
Mozambique	16.370	0.365	20.013	16.296	0.363	20.044
Uganda	14.909	0.174	18.644	14.979	0.176	18.615
Average	**15.955**	**0.235**	**19.855**	**15.956**	**0.234**	**19.875**

Table 4: Performance Comparison: Climate Model vs Baseline Model (PatchTST) across five countries

Metric	**UA-PatchTST**	**UA-ESTT**
Prediction Uncertainty (σ)		
Mean	2.6607	2.4499
Median	2.3710	2.0584
68% Intervals		
Coverage (%)	12.6	35.0
Avg. Width	5.3214	4.8997
95% Intervals		
Coverage (%)	28.8	53.1
Avg. Width	10.4299	9.6034

Table 5: Uncertainty Quantification Results

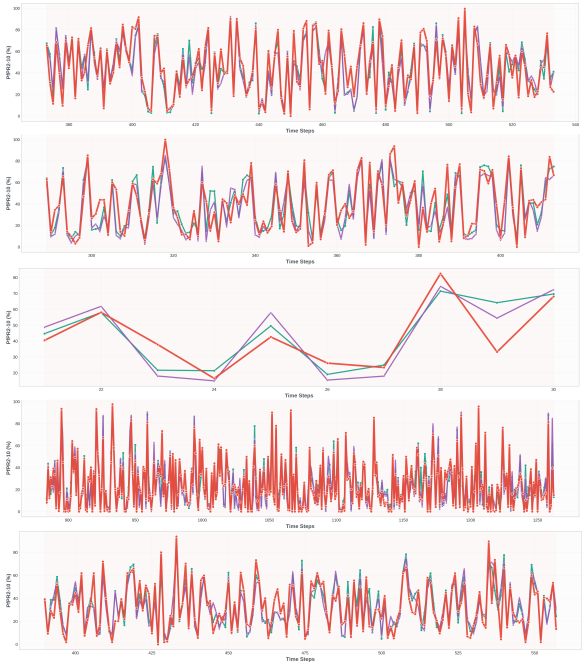


Figure 4: Future Forecast Details for Five Countries (Top to bottom: Mozambique, Nigeria, Congo, Tanzania, Uganda). Red: Actual values, Teal: Climate-enhanced model, Purple: Baseline model (ESTT)

Country-Specific Patterns: Congo exhibits the most substantial improvements with a 22.2% MAE reduction and a 2.0% R^2 enhancement, likely reflecting the strong climate-transmission relationships in this

equatorial high-burden setting. The climate-enhanced model effectively captures the pronounced seasonal patterns driven by temperature and precipitation cycles in the Congo Basin. Tanzania shows mixed results with a slight MAE increase of 2.2% but a small R^2 decrease (-1.15%), suggesting that climate variables may have less predictive power in this East African context or that the model struggles with Tanzania's more complex topographic and climatic heterogeneity. This finding highlights the importance of region-specific model calibration and the potential need for additional environmental variables in diverse geographic settings.

Nigeria, Mozambique, and Uganda demonstrate moderate but consistent improvements across some metrics, with MAE changes ranging from 0.6% improvement to 1.0% degradation. These results suggest that while climate variables provide valuable predictive information, their relative importance varies significantly across different transmission settings and geographic contexts.

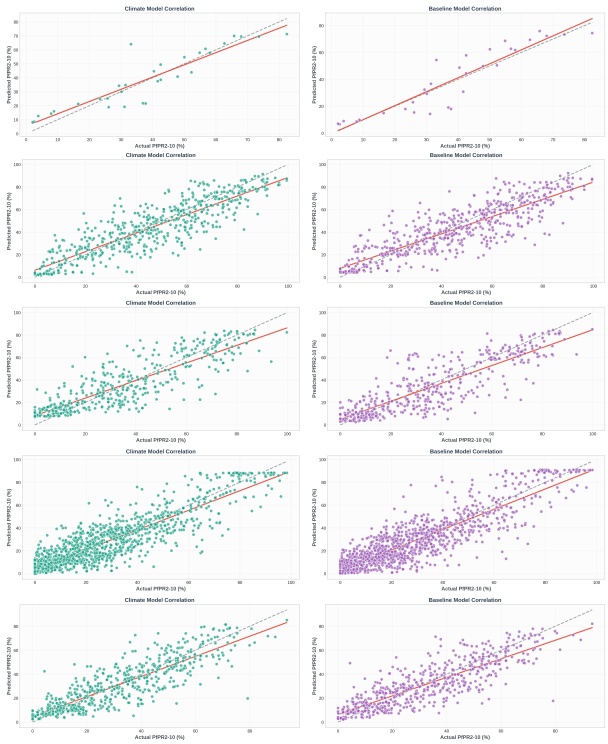


Figure 5: Climate and Baseline Model Correlation Analysis for Five Countries (Top to bottom: Congo, Mozambique, Nigeria, Tanzania, Uganda). Comparison between Climate-enhanced model and Baseline ESTT model performance

Uncertainty Estimation Analysis: The uncertainty analysis highlights distinct behaviors between UA-PatchTST and UA-ESTT. UA-PatchTST produces slightly higher average prediction uncertainty (mean $\sigma = 2.66$, median $= 2.37$) compared to UA-ESTT (mean $\sigma = 2.45$, median $= 2.06$), reflecting its tendency toward more conservative estimates. However, this comes at the cost of poor calibration, as UA-PatchTST captures far fewer true values within its intervals. In contrast, UA-ESTT achieves much higher coverage while maintaining tighter uncertainty ranges, indicating that its intervals are both narrower and more reliable. Overall, UA-ESTT strikes a better balance between uncertainty and calibration, making it the more dependable choice for reliable forecasting.

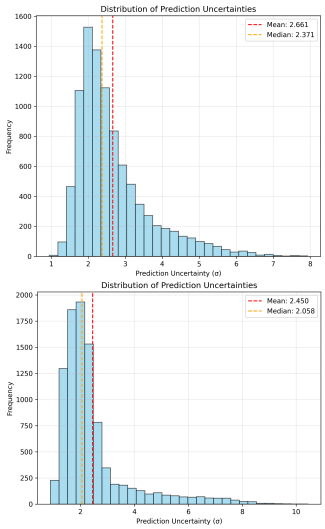


Figure 6: Uncertainty Evaluation(Top to bottom: UA-PatchTST, UA-ESTT)

6 Discussion

Accurate malaria forecasting has the potential to transform public health practice in Sub-Saharan Africa. Our findings demonstrate that neural network–based approaches consistently outperform traditional models, offering more reliable predictions to support evidence-based decision-making. By incorporating uncertainty quantification, these models go beyond point estimates, providing health practitioners with confidence-aware forecasts that enable risk-based planning and smarter allocation of limited resources. The spatial component enables identification of high-risk regions for targeted interventions, while temporal forecasting provides early warnings of outbreaks and seasonal surges, together supporting more proactive malaria control.

Future research could further explore integrating additional covariates such as mobility, land-use, and intervention coverage, as well as extending these models to real-time forecasting frameworks. Hybrid approaches that combine the strengths of spatiotemporal modeling and uncertainty quantification for more robust malaria early warning systems can be explored.

References

[1] F. Bbosa, R. Wesonga, J. Nabukenya, and P. Nabende, "Reliability of predictions using hybrid models: The case of malaria incidence rates in Uganda," *Journal of Health Informatics in Africa*, vol. 7, no. 2, pp. 29–46, 2020.

[2] A. Grosso, N. Hens, and S. Abrams, "An integrative review of the combined use of mathematical and statistical models for estimating malaria transmission parameters," *Malaria Journal*, vol. 24, no. 1, p. 173, 2025.

[3] H. H. Hussien, F. H. Eissa, and K. E. Awadalla, "Statistical methods for predicting malaria incidences using data from Sudan," *Malaria Research and Treatment*, vol. 2017, no. 1, p. 4205957, 2017.

[4] A. Koirala, M. Jha, S. Bodapati, A. Mishra, G. Chetty, P. K. Sahu, S. Mohanty, T. K. Padhan, J. Mattoo, and A. Hukkoo, "Deep learning for real-time malaria parasite detection and counting using YOLO-MP," *IEEE Access*, vol. 10, pp. 102157–102172, 2022.

[5] P. Martineau, S. K. Behera, M. Nonaka, R. Jayanthi, T. Ikeda, N. Minakawa, and Q. E. Mabunda, "Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning," *Frontiers in Public Health*, vol. 10, p. 962377, 2022.

[6] D. Mategula, J. Gichuki, K. I. Barnes, E. Giorgi, and D. J. Terlouw, "Advancing early warning systems for malaria: Progress, challenges, and future directions—A scoping review," *PLOS Global Public Health*, vol. 5, no. 5, p. e0003751, 2025.

[7] D. M. Megersa and X. S. Luo, "Effects of climate change on malaria risk to human health: A review," *Atmosphere*, vol. 16, no. 1, p. 71, 2025.

[8] S. R. Nayak, J. Nayak, S. Vimal, V. Arora, and U. Sinha, "An ensemble artificial intelligence-enabled MIoT for automated diagnosis of malaria parasite," *Expert Systems*, vol. 39, no. 4, p. e12906, 2022.

[9] World Health Organization, *World Malaria Report 2023*. World Health Organization, Geneva, 2023.

[10] K. K. Patro, A. J. Prakash, S. Madarapu, S. Ari, S. Routray, A. Mukherjee, and S. Mishra, "MalNet-DAF: Dual-attentive fusion deep learning model for malaria parasite classification," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[11] S. Rajab, J. Nakatumba-Nabende, and G. Marvin, "Interpretable machine learning models for predicting malaria," in *Proceedings of the 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, pp. 1–6, IEEE, 2023.

[12] A. Rajnarayanan, M. Kumar, and A. Tridane, "Analysis of a mathematical model for malaria using data-driven approach," *Scientific Reports*, vol. 15, no. 1, p. 27272, 2025.

[13] R. W. Snow, B. Sartorius, D. Kyalo, J. Maina, P. Amratia, C. W. Mundia, P. Bejon, and A. M. Noor, "The prevalence of *Plasmodium falciparum* in sub-Saharan Africa since 1900," *Nature*, vol. 550, no. 7677, pp. 515–518, 2017.

[14] D. Wang, S. Zhang, and L. Wang, "Deep epidemiological modeling by black-box knowledge distillation: An accurate deep learning model for COVID-19," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 15424–15430, 2021.

7 Appendix

In this section, we provide more details about Enhanced Spatio-Temporal Transformer and PatchTST based Spatio-Temporal Forecasting Architectures.

7.1 Nonlinear Spatiotemporal Operator Interpretation

We model malaria prevalence forecasting as learning a nonlinear spatiotemporal operator:

$$y(s, t+\Delta) \;=\; \mathcal{F}\Big(\{x(s, t-\tau)\}_{\tau=0}^{L-1},\; \phi(s)\Big),$$

where $x(s,t) \in \mathbb{R}^{d_{\text{temp}}}$ denotes temporal covariates (e.g., rainfall, temperature, previous prevalence) at location s and time t, and $\phi(s) \in \mathbb{R}^{d_{\text{spatial}}}$ denotes static spatial descriptors (e.g., geographic coordinates, elevation, population density). The operator $\mathcal{F}$ is parameterized by the proposed enhanced spatiotemporal transformer architecture.

7.1.1 Temporal Transformer as Data-Adaptive Temporal Convolution

In each self-attention block, given temporal embeddings $\{h_t\}_{t=1}^{L}$:

$$\tilde{h}_t = \sum_{u=1}^{L} \alpha_{t,u} V h_u, \quad \alpha_{t,u} = \text{softmax}\left(\frac{(Qh_t)(Kh_u)^\top}{\sqrt{d_k}} \right),$$

where Q, K, V are learned projections. Interpreting $\alpha_{t,\cdot}$ as a *data-dependent kernel* $K_\theta(t, u; \mathbf{x})$, we have:

$$\tilde{h}_t \;=\; \sum_{u=1}^{L} K_\theta(t, u; \mathbf{x})\,(Vh_u) \;\approx\; (K_\theta *_{\text{time}} h)_t.$$

Thus, multi-head attention performs a **nonlinear, input-conditioned convolution** over the temporal dimension. Stacking layers with feedforward networks yields a *Volterra-type operator*, which is well-suited for capturing delayed rainfall–vector–transmission pathways in malaria.

7.1.2 Spatial MLP as Location Embedding

Spatial features are mapped via:

$$S = \Psi_\theta(\phi(s)) \in \mathbb{R}^{d_{\text{model}}/2},$$

where Ψ_θ is a multi-layer perceptron with normalization, GELU activation, and dropout. This embedding encodes *static modifiers* such as endemicity, ecological factors, and connectivity.

7.1.3 Temporal–Spatial Fusion as Cross-Term Interaction

Fused representations are obtained via attention:

$$H_{\text{fused}} = \text{MHA}(H_{N_{\text{layers}}}, H_{N_{\text{layers}}}, H_{N_{\text{layers}}}),$$

followed by temporal pooling:

$$H_{\text{temp}} = \frac{1}{L} \sum_{t=1}^{L} H_{\text{fused}}[t, :].$$

The final prediction is:

$$\hat{y}(s, t+\Delta) = \Gamma_\theta\big([H_{\text{temp}}; S]\big),$$

where Γ_θ is the prediction MLP. Since H_{temp} already incorporates attention weights influenced by $\phi(s)$ through shared parameters, this effectively learns:

$$\hat{y}(s, t+\Delta) \;\approx\; \Gamma_\theta\left(\sum_{\tau=0}^{L-1} K_\theta(\tau; \{x(s, t-\cdot)\}, \phi(s))\, x(s, t-\tau) \right),$$

a **nonlinear spatiotemporal convolution** with spatially modulated, history-dependent kernels.

7.1.4 Approximation Capacity

Transformers with GELU activations and MLPs are universal sequence-to-scalar approximators on compact sets. Hence, the architecture can approximate continuous operators such as Volterra or Wiener functionals, which capture lagged climate effects in malaria dynamics. Learnable positional embeddings ensure temporal order awareness, preventing permutation invariance in time.

7.1.5 Stability Properties

With norm-first residual blocks, the per-layer Lipschitz constant satisfies:

$$\mathrm{Lip}(f_\theta) \lesssim \prod_{\ell=1}^{N} \big(1 + L_{\mathrm{attn}}^{(\ell)}\big)\big(1 + L_{\mathrm{ffn}}^{(\ell)}\big),$$

where $L_{\mathrm{attn}}^{(\ell)}$ and $L_{\mathrm{ffn}}^{(\ell)}$ are layer-specific constants. Layer normalization keeps these factors moderate, while dropout reduces overfitting. Temporal averaging (pooling) is a contractive step (operator norm ≤ 1), further stabilizing predictions.

7.1.6 Domain Alignment

Malaria prevalence depends on multi-month, nonlinear lags (rainfall $\rightarrow$ mosquito breeding $\rightarrow$ transmission). Data-adaptive kernels capture *when* and *how strongly* to weight past months. Spatial embeddings inject location-specific priors, allowing kernel shapes to vary across regions.

7.2 Patch-Based Time Series Decomposition

The core insight of PatchTST is that time series can be effectively decomposed into non-overlapping or overlapping patches that capture local temporal patterns more efficiently than full-sequence attention mechanisms.

7.2.1 Patch Creation and Embedding

For a given time series $X \in \mathbb{R}^{T \times d}$, we create patches of length P with stride S:

$$\mathrm{Patch}_j = X[j \cdot S : j \cdot S + P, :], \quad j = 0, 1, \ldots, \lfloor \frac{T - P}{S} \rfloor$$

Each patch is then flattened and linearly projected to the model dimension:

$$E_j = \mathrm{Linear}(\mathrm{Flatten}(\mathrm{Patch}_j)) \in \mathbb{R}^{d_{model}}$$

where the linear projection transforms patches from $\mathbb{R}^{P \times d}$ to $\mathbb{R}^{d_{model}}$. This patch-based approach reduces the sequence length from T to $N_p = \lfloor \frac{T-P}{S} \rfloor + 1$ patches, significantly improving computational efficiency while preserving local temporal structure.

7.2.2 Positional Encoding for Patches

To maintain temporal ordering information, we add learnable positional encodings to patch embeddings:

$$\tilde{E}_j = E_j + \mathrm{PE}_j$$

where $\mathrm{PE}_j \in \mathbb{R}^{d_{model}}$ represents learnable position embeddings that encode the temporal position of each patch within the sequence.

7.3 Channel-Independent Processing Architecture

A fundamental innovation of PatchTST is channel-independent processing, where each feature dimension is processed by a separate transformer encoder. This approach is particularly beneficial for malaria forecasting where epidemiological variables (age groups, examination counts) and climate variables (temperature, precipitation) exhibit different temporal dynamics and scaling properties.

7.3.1 Individual Channel Transformers

For each feature channel $c \in \{1, 2, \ldots, d\}$, we extract the corresponding data:

$$X^{(c)} = X[:, c] \in \mathbb{R}^T$$

Channel-specific patches are created and processed by dedicated transformer encoders:

$$H^{(c)} = \mathrm{TransformerEncoder}_c(\mathrm{PatchEmbed}(X^{(c)}))$$

where each $\mathrm{TransformerEncoder}_c$ consists of L transformer layers with multi-head self-attention:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

$$\mathrm{MultiHead}(X) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O$$

where $\mathrm{head}_i = \mathrm{Attention}(XW_i^Q, XW_i^K, XW_i^V)$.

7.3.2 Channel Aggregation and Global Representation

After independent processing, channel representations are aggregated to form a unified temporal representation:

$$H_{temporal} = \mathrm{ChannelAggregation}([H^{(1)}; H^{(2)}; \ldots; H^{(d)}])$$

where $[\cdot;\cdot]$ denotes concatenation and the aggregation function learns optimal channel combination:

$$H_{temporal} = \mathrm{Linear}(\mathrm{Concat}(H^{(1)}, \ldots, H^{(d)})) \in \mathbb{R}^{d_{model}}$$

7.4 Spatial Feature Integration

To account for the significant spatial heterogeneity in geographic and climatic conditions, our architecture explicitly models location-specific attributes. This is achieved through a dedicated sub-network that processes geospatial features to learn a distinct embedding for each location, thereby conditioning the forecast on its unique regional context.

7.4.1 Spatial Feature Projection

Spatial features $s \in \mathbb{R}^{d_s}$ including coordinates and location-specific climate statistics undergo projection and nonlinear transformation:

$$S_{proj} = \text{Linear}(s) \in \mathbb{R}^{d_{model}}$$

$$S_{processed} = \text{SpatialMLP}(\text{LayerNorm}(S_{proj}))$$

where the spatial MLP consists of multiple layers with GELU activations and dropout regularization:

$$\text{SpatialMLP}(x) = \text{Dropout}(\text{GELU}(\text{Linear}(\text{LayerNorm}(x))))$$

7.5 Attention-Based Spatiotemporal Fusion

The integration of temporal and spatial representations employs cross-attention mechanisms that learn optimal feature interactions for malaria prediction.

7.5.1 Cross-Attention Fusion

Temporal features serve as queries while spatial features provide additional context through key-value pairs:

$$H_{fused}, \alpha = \text{MultiHeadAttention}(H_{temporal}, H_{temporal}, H_{temporal})$$

where the attention weights α provide interpretability regarding which temporal periods contribute most to predictions.

7.5.2 Final Prediction Network

The fused representation combines temporal and spatial information for prevalence prediction:

$$\hat{y} = \text{PredictionMLP}([H_{fused}; S_{processed}])$$

The prediction network employs multiple layers with residual connections and layer normalization:

$$\begin{aligned}&\text{PredictionMLP}(x) = \\ &\text{Linear}(\text{Dropout}(\text{GELU}(\text{LayerNorm}(\text{Linear}(x))))).\end{aligned}$$

7.6 Training Methodology and Optimization

7.6.1 Loss Function and Regularization

Training employs Smooth L1 Loss for robustness against outliers common in epidemiological data:

$$\mathcal{L}(y, \hat{y}) = \begin{cases} 0.5(y - \hat{y})^2 & \text{if } |y - \hat{y}| < 1 \\ |y - \hat{y}| - 0.5 & \text{otherwise} \end{cases}$$

L2 regularization is applied to prevent overfitting:

$$\mathcal{L}_{total} = \mathcal{L}(y, \hat{y}) + \lambda \sum_{\theta} \theta^2$$

7.6.2 Advanced Optimization Strategy

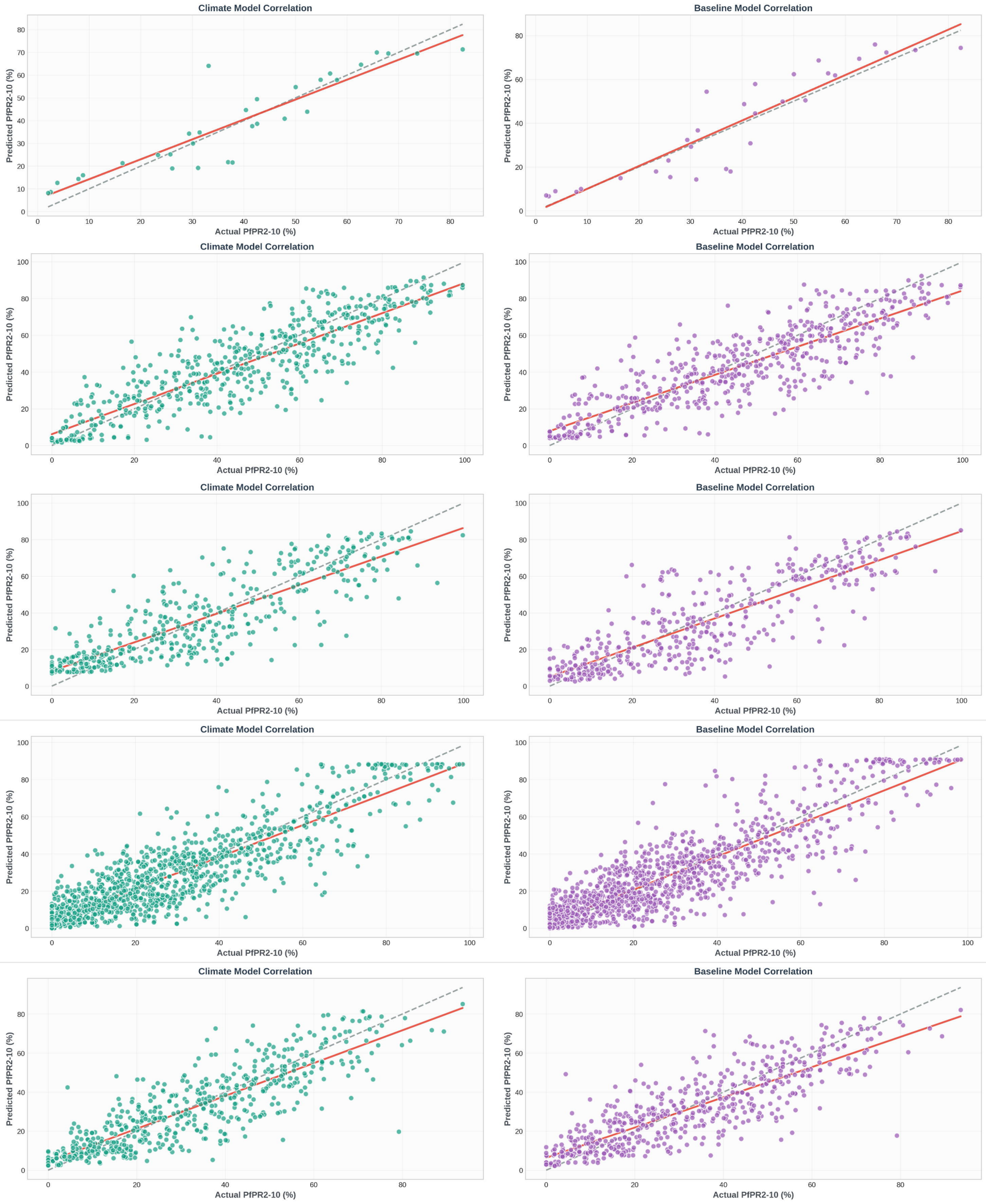
The optimization employs AdamW with cosine annealing warm restarts to escape local minima:

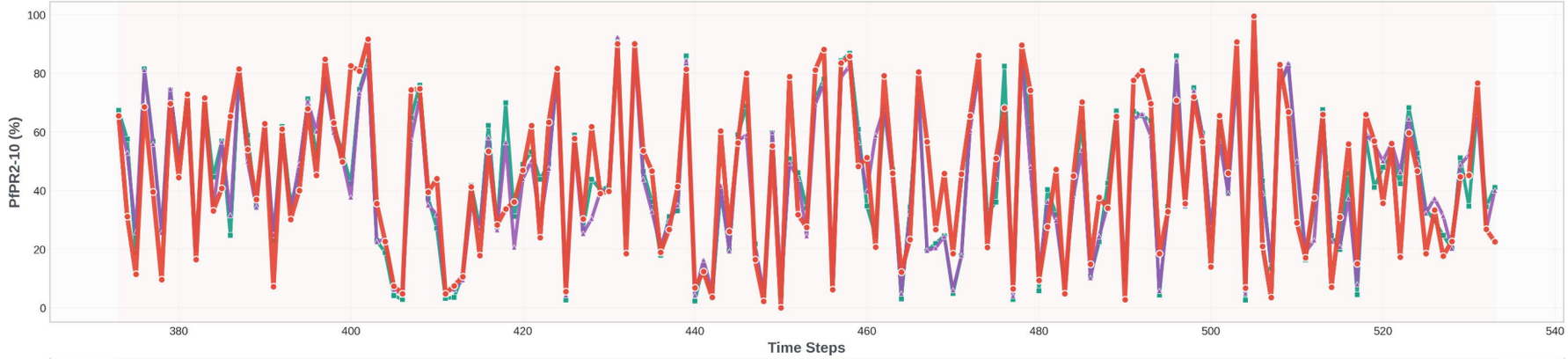
$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right)$$

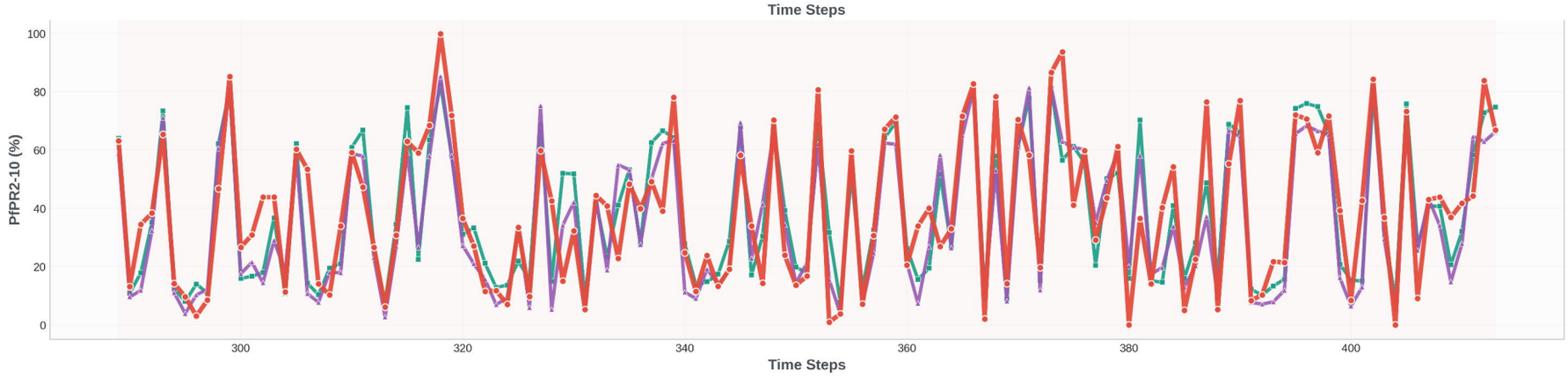
where T_{cur} represents the current epoch within restart cycle T_i, and $T_i = T_0 \cdot T_{mult}^i$ grows exponentially to allow for longer exploration periods.

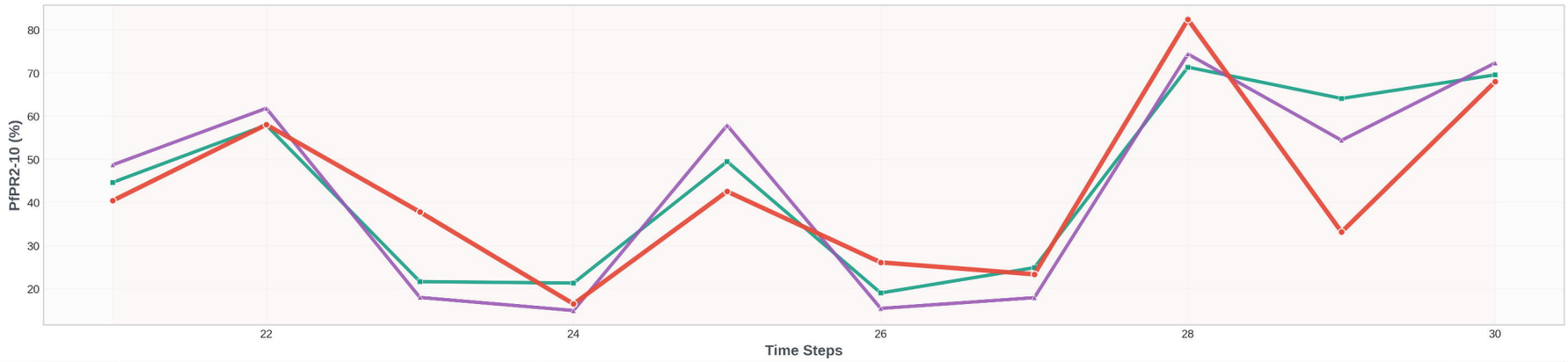
Gradient clipping with maximum norm constraint ensures training stability:

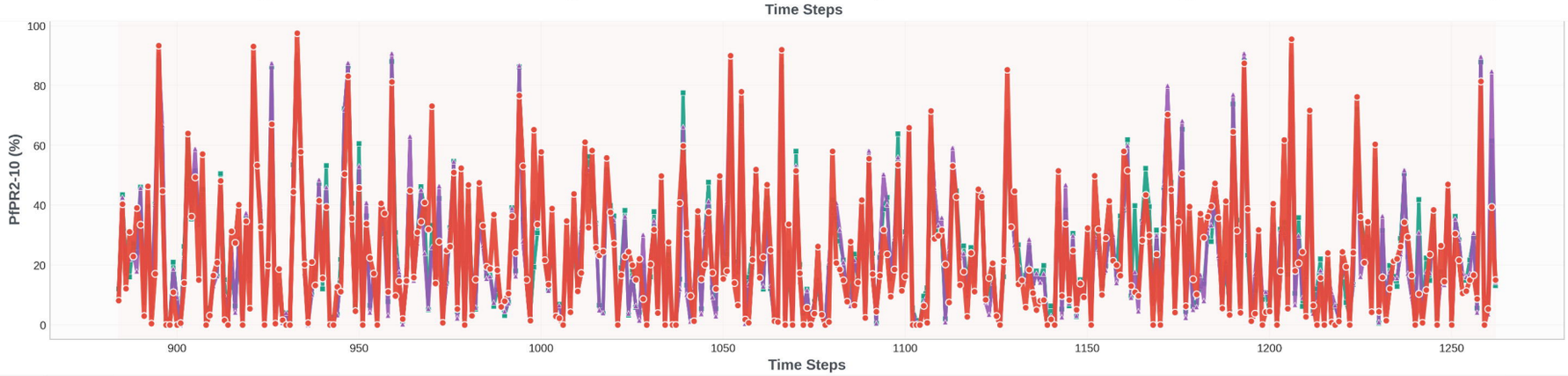
$$\text{if } \|\nabla\theta\| > \text{clip_norm}: \quad \nabla\theta \leftarrow \frac{\text{clip_norm} \cdot \nabla\theta}{\|\nabla\theta\|}$$

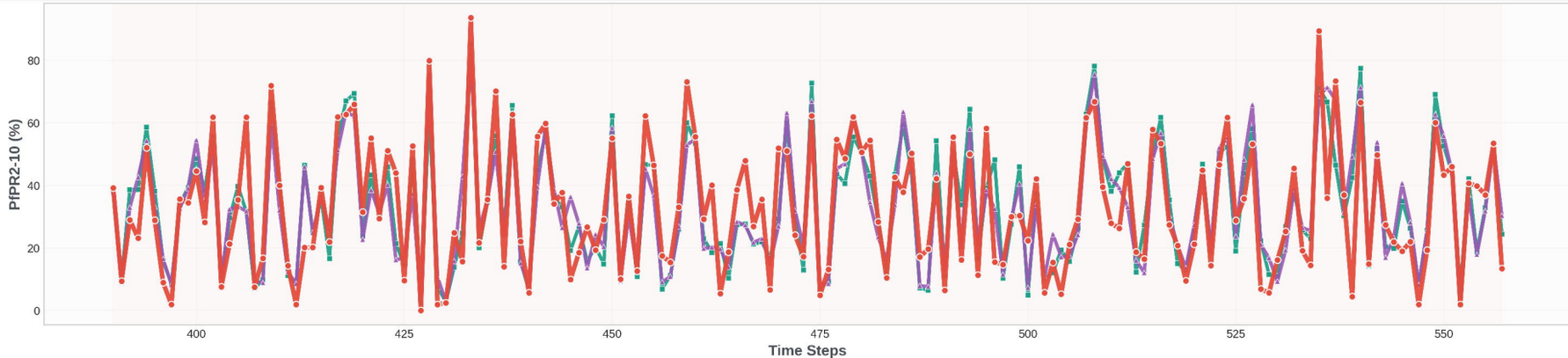

Climate Model Correlation
Baseline Model Correlation
Predicted PfPR2-10 (%)
Actual PfPR2-10 (%)

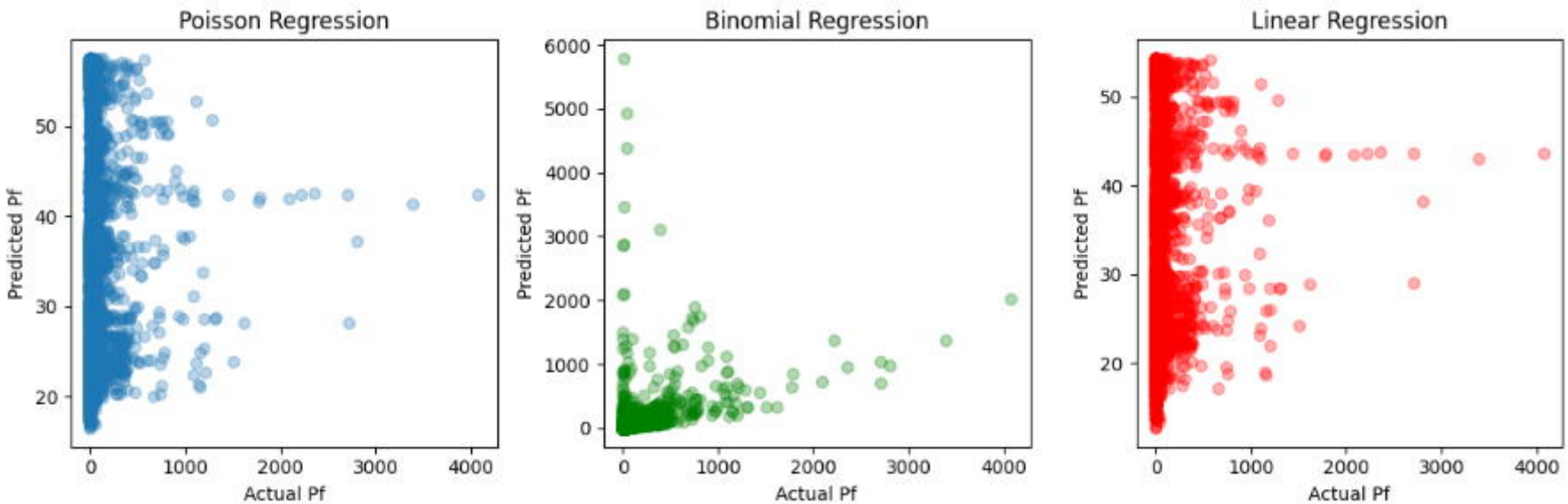

PfPR2-10 (%)
Time Steps

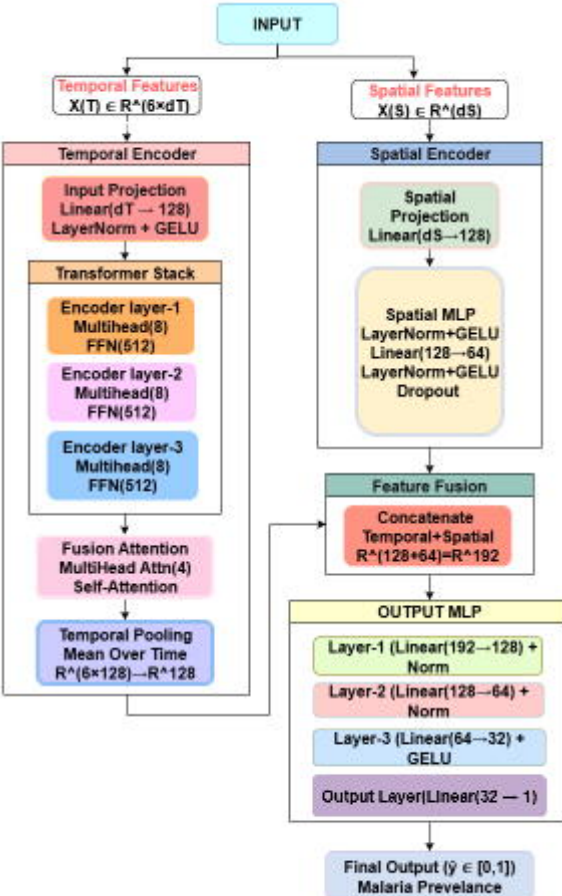

PfPR2-10 (%)
Time Steps

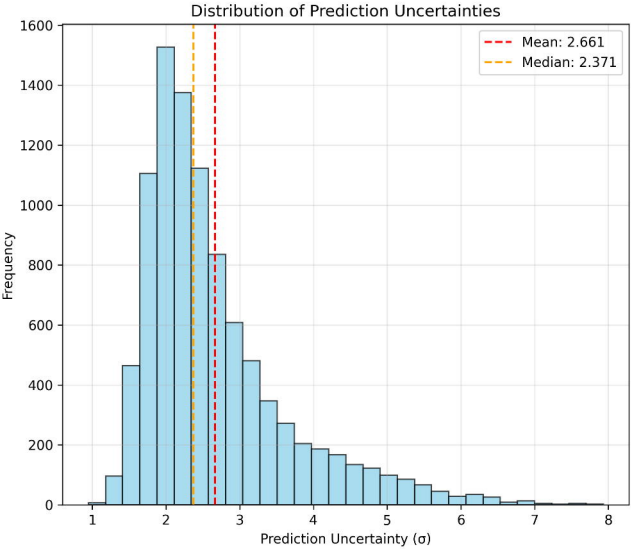

PfPR2-10 (%)
Time Steps

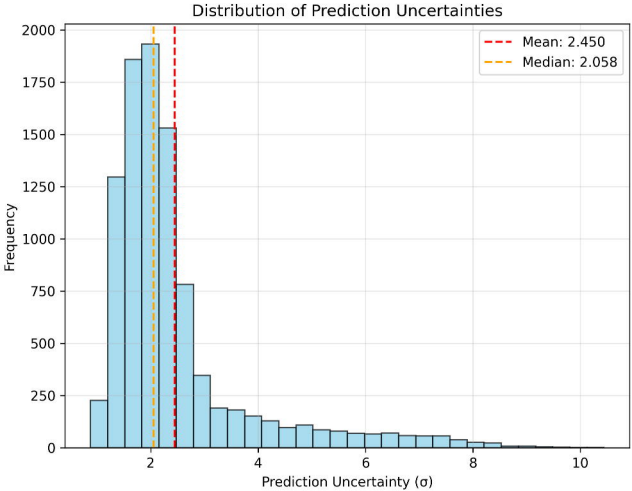

PfPR2-10 (%)
Time Steps

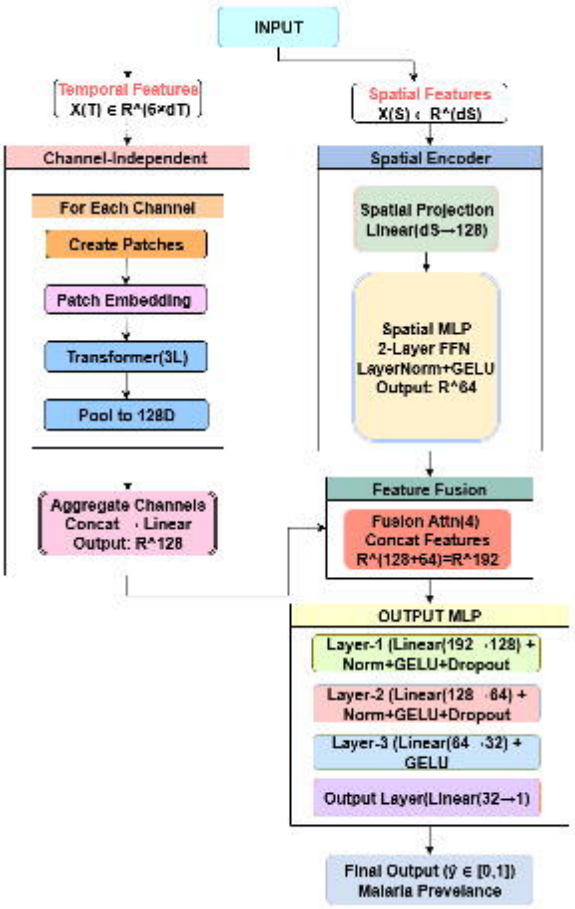

PfPR2-10 (%)
Time Steps


Poisson Regression
Predicted Pf
Actual Pf
Binomial Regression
Predicted Pf
Actual Pf
Linear Regression
Predicted Pf
Actual Pf


INPUT
Temporal Features
X(T) ∈ R^(6×dT)
Spatial Features
X(S) ∈ R^(dS)
Temporal Encoder
Input Projection
Linear(dT → 128)
LayerNorm + GELU
Transformer Stack
Encoder layer-1
Multihead(8)
FFN(512)
Encoder layer-2
Multihead(8)
FFN(512)
Encoder layer-3
Multihead(8)
FFN(512)
Fusion Attention
MultiHead Attn(4)
Self-Attention
Temporal Pooling
Mean Over Time
R^(6×128)→R^128
Spatial Encoder
Spatial
Projection
Linear(dS→128)
Spatial MLP
LayerNorm+GELU
Linear(128→64)
LayerNorm+GELU
Dropout
Feature Fusion
Concatenate
Temporal+Spatial
R^(128+64)=R^192
OUTPUT MLP
Layer-1 (Linear(192→128) + Norm
Layer-2 (Linear(128→64) + Norm
Layer-3 (Linear(64→32) + GELU
Output Layer(Linear(32 — 1)
Final Output (ŷ ∈ [0,1])
Malaria Prevelance


Distribution of Prediction Uncertainties
Mean: 2.661
Median: 2.371
Frequency
Prediction Uncertainty (σ)
0
200
400
600
800
1000
1200
1400
1600
1
2
3
4
5
6
7
8


Distribution of Prediction Uncertainties
Mean: 2.450
Median: 2.058
Frequency
Prediction Uncertainty (σ)
0
250
500
750
1000
1250
1500
1750
2000
2
4
6
8
10


INPUT
Temporal Features
X(T) ∈ R^(6×dT)
Spatial Features
X(S) ∈ R^(dS)
Channel-Independent
For Each Channel
Create Patches
Patch Embedding
Transformer(3L)
Pool to 128D
Aggregate Channels
Concat → Linear
Output: R^128
Spatial Encoder
Spatial Projection
Linear(dS→128)
Spatial MLP
2-Layer FFN
LayerNorm+GELU
Output: R^64
Feature Fusion
Fusion Attn(4)
Concat Features
R^(128+64)=R^192
OUTPUT MLP
Layer-1 (Linear(192→128) +
Norm+GELU+Dropout
Layer-2 (Linear(128→64) +
Norm+GELU+Dropout
Layer-3 (Linear(64→32) +
GELU
Output Layer(Linear(32→1)
Final Output (ŷ ∈ [0,1])
Malaria Prevalence