Zero-Shot Dengue Forecasting with LLM-based Time Series Foundation Models

Pratik S. Machchar
Research Scholar
Gujarat Technological University,
Ahmedabad, India

Purvi N. Ramanuj, PhD
Associate Professor
Vishwakarma Government Engineering College,
Ahmedabad, India

ABSTRACT

Worldwide, approximately 390 million dengue infections occur annually. Health systems in the tropics struggle to prepare for outbreaks, even when there is forewarning of impending surges that threaten to overrun hospitals. This paper explores whether recent large language model (LLM)-based time-series foundation models can address this issue, or whether they exhibit the same limitations as classical models. This study systematically benchmarks four zero-shot foundation models—Chronos-2 (Amazon), Google TimesFM 2.5, Salesforce MOIRAI, and Tsinghua Sundial—against an auto-tuned ARIMA baseline using epidemiological data from San Juan (Puerto Rico) and Iquitos (Peru) with climate covariates (temperature, dew point, precipitation, and vegetation indices). Evaluated over a 12-week forecasting horizon using exact metrics from verified experimental runs, model performance diverged across cities: Chronos-2 Multivariate achieved top accuracy in San Juan (RMSE 3.7604, MAE 3.1203), outperforming ARIMA (RMSE 3.9082), closely followed by Google TimesFM 2.5 (RMSE 4.0705) and Salesforce MOIRAI (RMSE 4.6054, MAE 4.1959). In Iquitos, Tsinghua Sundial flow-matching attained the minimal error (RMSE 2.5039, MAE 2.2139), closely followed by Google TimesFM 2.5 (RMSE 2.5487, MAE 2.2428) and Salesforce MOIRAI (RMSE 2.8459, MAE 2.5089). Beyond providing a model ranking, this study articulates the architectural reasons behind these performance differences. The findings conclude that model selection must be guided by the target city's epidemic amplitude, the availability of computational resources, and the requirement for calibrated uncertainty estimates.

General Terms

Dengue Forecasting

Keywords

Forecasting Dengue Fever, Chronos-2, Salesforce MOIRAI, TimesFM, Tsinghua Sundial, time series forecasting, foundation model, forecasting using LLMs, multivariate forecasting, epidemiological forecasting, outbreak monitoring, San Juan, Iquitos, climate-driven disease forecasting

1. INTRODUCTION

The challenge of dengue forecasting has been present for over two decades now, and despite extensive research, it remains challenging. Official estimates range from 100 million to 390 million cases per year—this uncertainty alone highlights the difficulty of tracking and forecasting the disease [1]. At least 129 nations harbor endemic dengue [2], costs exceed $39 billion annually [3], and 2023 saw the largest number of cases ever reported in the Americas [4]. Nonetheless, most public health agencies continue to use ARIMA models based on mathematical assumptions made in the 1970s.

The difficulty in forecasting the disease arises from the complex interplay of several factors. Ambient temperature controls the pace at which dengue virus cycles through its reproductive cycle within the mosquito host, with replication rapidly declining below 26–29°C [5]. Rainfall and humidity define the number of available breeding habitats as well as mosquito longevity [6]. NDVI of vegetation cover, obtained from satellite imagery, defines the intricate microclimate required for vector mosquitoes living in an urban habitat [7]. These indicators do not impact weekly case numbers directly but through a cascade of four-to-eight-week delays, starting from precipitation, going through larval development, emergence of adult mosquitoes, blood feeding, infection period and symptoms appearance. Any forecasting algorithm that disregards these temporal lags cannot perform reliably, regardless of its architectural sophistication.

The classical ARIMA and SARIMA models have several disadvantages in this context: they assume a linear system and stationary processes, which do not withstand the impact of actual outbreaks or long-term dynamics due to serotype replacement [8]. On the other hand, gradient boosting algorithms, and generally speaking, machine learning algorithms, overcame the limitation of the assumption of linearity and demonstrated positive results when manually engineering lagged climatic parameters [9, 10]. LSTM algorithms outperformed this approach, as they automatically detected temporal dependencies based on sequential outbreaks without requiring explicit lag specification [11]. However, all of these models share one common limitation: they require sufficient local history data (up to several years). In a resource-limited health system serving a network of endemic districts, that is a serious obstacle. There is also a largely unaddressed issue that almost none of these methods are able to handle multivariate data natively, despite the fact that alert-threshold decisions in public health genuinely require them [12].

This context motivates the exploration of zero-shot time series foundation models. The core premise is appealing: pre-train a large neural architecture on heterogeneous time-series corpora, then deploy it directly to unseen epidemiological surveillance series without city-specific fine-tuning. Prominent foundation model paradigms include group-attention transformers (Chronos-2 [13]), continuous patch-decoder networks (Google TimesFM 2.5 [14]), any-variate masked encoder transformers (Salesforce MOIRAI [15]), and continuous flow-matching vector field models (Tsinghua Sundial [16]). While early benchmarks on generic datasets show promise, a disciplined, controlled evaluation spanning these distinct zero-shot foundation architectures specifically on dengue case data—incorporating climate covariates across epidemiologically distinct cities such as San Juan and Iquitos—remains unaddressed. This work aims to bridge that gap. Contributions include:

(i) A comparative benchmark evaluating four zero-shot foundation model paradigms—Chronos-2 Multivariate, Google TimesFM 2.5, Salesforce MOIRAI, and Tsinghua Sundial—against an auto-tuned ARIMA baseline across epidemiologically distinct cities (San Juan and Iquitos).

(ii) Quantification of multivariate climate conditioning (temperature, dew point, precipitation, NDVI) in Chronos-2 relative to univariate patch baselines.

(iii) Sensitivity analysis of zero-shot model failure modes under extreme epidemic peak amplitudes.

Section 2 reviews background literature. Section 3 details the four foundation model architectures. Section 4 covers the data and experimental setup. Section 5 reports results and discussion. Section 6 identifies open questions. Section 7 concludes.

2. BACKGROUND AND RELATED WORK

2.1 Classical Statistical Methods

The use of time-series analysis to forecast dengue fever has been around since at least the beginning of the 2000s. ARIMA-type models [17] have been used in Thailand [18], Brazil [19], and Puerto Rico [20]. These models perform adequately in the periods between epidemics when the number of weekly cases changes gradually. What becomes evident from the comparative literature on the issue [21] is that ARIMA-type models tend to underestimate epidemic peak levels; these models will tend to flatten out the peak instead of predicting it. It should be noted that this is not due to a lack of adequate tuning but is a direct result of the assumption of linearity embedded in such models.

2.2 Machine Learning and Deep Learning

Transitioning to the use of gradient boosting machines and support vector regression models proved effective because they do not rely on linear assumptions, as demonstrated by Buczak et al. [9] in the Philippines. This conclusion was further supported by Guo et al. [10], who demonstrated that cumulative rainfall with a lag of four to six weeks provided an important portion of weekly case number variance unexplained by any autoregressive model. Gupta et al. [22] demonstrated similar benefits by comparing deep learning ensembles that also included socioeconomic variables with single-model machine learning baselines, which is consistent with state-level deep learning assessments in Brazil incorporating climate lags and spatial dependencies [23]. The trend observed in all three works is clear: nonlinear feature interaction and time lags improve model performance; however, traditional model limitations persist.

Subsequently, Long Short-Term Memory (LSTM) networks made learning temporal dependencies feasible [24]. The hybrid approach by Zhang [25], fitting ARIMA for the trend component and applying LSTM to residuals, was an influential design for such problems due to its interpretability in the trend component alongside flexible modelling of residuals. Attention mechanisms, such as those used in PatchTST [26], provided improved stability for long-term forecasting without the gradient issues of recurrent nets. Nonetheless, none of this solves the core problem of dependence on data: LSTM and transformers trained on outbreaks in one city do not generalise well to other cities, and much of the endemic areas around the world lack sufficient historical labeled data.

2.3 Time Series Foundation Models

The concept of foundation models —pre-training on scale, deployment in zero-shot fashion— started from NLP and computer vision domains before moving on to time series forecasting. Chronos [27], created by Amazon, stands out as an early milestone: it discretises numerical values into tokens, uses a transformer-based encoder-decoder architecture (T5) to pre-train a model on 84 billion data points of heterogeneous nature, and performs competitively in zero-shot settings when compared with fine-tuned task-oriented models on 42 benchmarks. The possibility that predictions based on Chronos could be competitive even on unseen epidemic-related data prompted to conduct the thorough evaluation reported in this work.

Google's TimesFM 2.5 [14] handles numerical series by mapping continuous patch embeddings into a decoder-only transformer architecture with direct regression heads, eliminating value-quantisation noise. Salesforce MOIRAI [15] introduces an any-variate masked encoder framework, utilizing multi-patch projection layers and variate-agnostic self-attention trained on the 27-billion-observation LOTSA dataset. Most recently, Tsinghua Sundial [16] introduced a continuous flow-matching approach to time-series foundation modelling, predicting vector-field velocity trajectories via probability flow Ordinary Differential Equations (ODEs). This enables Sundial to capture non-Gaussian, multimodal time-series probability distributions without token discretisation or fixed point regression heads.

Additional architectures of interest include: Lag-Llama [28], which utilised the Llama architecture to perform probabilistic forecasting with univariate data through lag-based tokenization; AutoTimes [29], which experimented with text-based autoregressive prompting on timeseries data; and EpiLLM [30], which focused specifically on epidemiological timeseries. To facilitate zero-shot disease forecasting without historical data, simulation-grounded foundation models like Mantis [31] have been introduced. Additionally, systematic benchmarking efforts have evaluated how general-purpose time-series foundation models perform in zero-shot settings relative to traditional epidemiological models and policy evaluation scenarios [32]. Decomposition transformers such as Autoformer [33], FEDformer [34], and TimesNet [35] have also been influential in the forecasting community, though they do not fit the foundation model framework and were not included in this comparison.

In a recent paper that is directly relevant to this contribution. Guo et al. [36] ran a systematic multi-model dengue comparison in Guangzhou and found that adding climate features to tree-based models reduced the RMSE by 18–24% over univariate baselines. To extend this line of inquiry into foundation models, the native multivariate covariate support added in Chronos-2 relative to the original Chronos is the single most important architectural development enabling a fair test of climate inputs.

3. MODEL ARCHITECTURES

Four LLM models are evaluated: Chronos-2, Salesforce MOIRAI, Google TimesFM 2.5, and Tsinghua Sundial because together they span four fundamental architectural approaches in time-series foundation modelling: group-attention multivariate encoders, any-variate masked encoder transformers, continuous patch decoders, and continuous flow-matching ODE vector field models. By testing all four using the same data and evaluation protocol, the structural features essential for zero-shot dengue forecasting are determined.

3.1 Chronos-2

Chronos-2 [13] is the successor to the original Chronos model [27], both developed by Amazon Science. The original Chronos introduced a univariate framework that discretised time series values into 4096 tokens and processed them through a T5 encoder-decoder architecture pre-trained on 84 billion data points. While it demonstrated competitive zero-shot performance on 42 benchmarks, its strictly univariate design prevented the use of exogenous covariates.
Chronos-2 addresses this limitation by adopting a fundamentally different architecture: an **encoder-only** transformer with a novel **group attention** mechanism [13]. Instead of quantising values into discrete tokens, Chronos-2 normalises each input series and divides it into non-overlapping patches that are mapped to continuous high-dimensional embeddings via a residual network. Each transformer block alternates between two attention layers:

Time attention: aggregates information across patches within a single time series.

Group attention: shares information across all series within a user-defined group (e.g., the dengue target together with its five climate covariates).

This group attention mechanism is what enables native multivariate and covariate-informed forecasting without any task-specific training. The model outputs direct multi-step quantile forecasts (21 quantiles from 0.01 to 0.99), eliminating the autoregressive token-by-token sampling of the original Chronos.
Chronos-2 is available in 28M and 120M parameter variants and supports context lengths of up to 8192 tokens. In this dengue experiments the 5-feature multivariate version produced RMSE = 3.7604 in San Juan and RMSE = 3.7150 in Iquitos.

3.2 Salesforce MOIRAI

Salesforce MOIRAI [15] (Masked Encoder Any-variate Transformer) is a universal time-series foundation model pre-trained on the Large-scale Open Time Series Archive (LOTSA), comprising 27 billion observations across diverse domains. MOIRAI addresses the heterogeneity of time-series data by employing an Any-variate attention mechanism that handles arbitrary numbers of input channels and multi-patch projection layers that process variable patch lengths (e.g., 8, 16, 32, 64, 128).
Unlike classical autoregressive models or rigid single-patch decoders, MOIRAI models multivariate target and covariate series as flattened sequences of patch tokens with masked self-attention. The network outputs a mixture of Student's t-distributions over future time steps, enabling native probabilistic forecasting without value-quantisation noise. In this benchmark, the official zero-shot MOIRAI model was deployed (`Salesforce/moirai-1.0-R-small`) over a 52-week context window and 12-week prediction horizon without task-specific fine-tuning.

3.3 TimesFM 2.5

TimesFM 2.5 [14] (`google/timesfm-2.5-200m-pytorch`, 200M parameters) follows the continuous-value patch approach. Windowed inputs are mapped to dense embeddings and then sent through a transformer-based decoder network; point predictions are generated by the regression head, and optional quantile range P10-P90 estimates may be produced. Since the encoding process does not involve discretization, no quantisation error exists. This is a significant benefit, especially when dealing with single-digit cases per week, because a vocabulary size of 4096 causes quantisation noise that overwhelms the signal. Training was done on approximately 100 billion time steps of retail, energy, web, and traffic data. The 200M model is easily run without a GPU, which is a practical consideration for health agencies in low-resource locations.
In Iquitos, TimesFM 2.5 achieved RMSE = 2.5487 (MAE 2.2428), demonstrating strong zero-shot precision on Iquitos's weekly case series. San Juan delivered RMSE = 4.0705 (MAE 3.4587). The performance difference between cities stems from distribution amplitude mismatch, as pre-training datasets predominantly feature non-epidemic time-series signatures.

3.4 Tsinghua Sundial

Tsinghua Sundial [16] introduces a generative continuous flow-matching architecture for time-series forecasting. Unlike discretised language models (Chronos T5) or direct point regression heads (TimesFM, MOIRAI), Sundial formulates forecasting as learning a continuous probability flow Ordinary Differential Equation (ODE) vector field over normalized time-series trajectories.
During pre-training across heterogeneous multi-domain time series, Sundial learns a neural velocity field $v_\theta(x_t, t)$ that transforms a simple Gaussian prior $p_0(x)$ into the target predictive distribution $p_1(x)$ via continuous flow matching. At inference time, multi-step probabilistic trajectory samples are drawn by solving the probability flow ODE using numerical integration step solvers.
Sundial supports context windows up to 2880 steps and provides native probabilistic output without autoregressive sampling noise. In the zero-shot dengue benchmarks, Sundial demonstrated high precision in Iquitos's endemic series, achieving the top zero-shot RMSE of 2.5039 (MAE 2.2139) in Iquitos.

3.5 Comparison of the Four Foundation Models

Table 1 summarizes the core architectural differences among the four zero-shot foundation models evaluated in this study.

4. DATA AND EXPERIMENTAL SETUP

4.1 Study Cities

The selection of San Juan and Iquitos was deliberate. They lie on different ends of the dynamic spectrum that the surveillance program for dengue actually encounters.
San Juan, Puerto Rico (18.5°N, 66.1°W) exhibits dengue as seasonal epidemic waves but with occasional outbreaks of greater severity. The baseline —number of cases during the low-transmission season of winter —is not a problem. The issue lies with the years when the number of weekly cases went above 460, such as in 1994, 2010, and 2012. Amplitude variability makes forecasting models perform qualitatively differently depending on whether they are predicting an outbreak or just an increase in cases during the rainy season. Johansson et al. have set up an open forecasting challenge for San Juan [12]; however, submissions even from groups with solid epidemiological background were unable to reliably predict the outbreak magnitude [4].
Iquitos, Peru (3.7°S, 73.2°W) lies in the Peruvian Amazon region with high temperatures and humidities all throughout the year. Weekly case numbers vary but are low; test period maxima did not exceed 45 cases per week as identified in the analysis. Seasonality is present but not so much as compared to that of San Juan. In contrast, Iquitos serves as a testbed for evaluating whether a forecasting framework can capture the variations in low-to-moderate case numbers, despite occasional spikes beyond typical seasonal levels. This problem might be a better reflection of the reality facing South and Southeast Asian districts, than the epidemic fluctuations of San Juan. Epidemiological background on both cities is reviewed in [7, 22].

Table 1. Architecture Comparison of the Four Evaluated Foundation Models

Attribute	Chronos-2	Salesforce MOIRAI	TimesFM 2.5	Tsinghua Sundial
Architecture	Encoder-only (group attn)	Masked Encoder	Dec-only dense	Flow-Matching ODE
Input representation	Continuous patches	Multi-patch projection	Continuous patches	Continuous normalised
Probabilistic output	Quantiles (21 levels)	Mixture (Student's t)	Quantiles (P10–P90)	Flow Trajectories
Multivariate covariates	Native (group attention)	Native (Any-variate)	Limited	Channel-Independent
Parameter range	28M / 120M	14M / 91M / 311M	200M	$\sim$128M
Context window	Up to 8192 steps	Up to 512 steps	Up to 16K steps	Up to 2880 steps
Pre-training scale	Synthetic multivariate	LOTSA (27B observations)	$\sim$100B points	Multi-domain corpus
Zero-shot deployment	Yes	Yes	Yes	Yes
GPU required	Recommended	Recommended	No (CPU feasible)	Recommended

4.2 Dataset

The benchmark data was obtained from the DengAI competition dataset, publicly available via DrivenData. This dataset contains weekly observations from San Juan between 1990 and 2010, and from Iquitos between 2000 and 2010. Each record contains the number of weekly dengue cases confirmed, as well as all possible weather variables and vegetation variables: several temperature variables, dew point, relative humidity, total precipitation, and four separate NDVI measurements based on MODIS satellite imagery.
In the multivariate models, however, not all available variables were utilized. Permutation importance testing on a held-out subset of data that five features drove the majority of the predictive signal: average temperature, dew point, relative humidity, total precipitation, and NDVI in the northeast quadrant. All of these variables have been shown to play a role in the known mechanistic pathways between weather and dengue infection [36, 6]. Using all fifteen-plus possible variables led to suboptimal results, as discussed in the results section.

4.3 Experimental Protocol

Multiple scenarios were analysed, considering changes in context size, forecasting period, covariates, and model version. The three important ones are highlighted here:

Exp-2 (C-52, H-12): 52-week context, 12-week horizon, 80/20 chronological split. Primary cross-model comparison.

Exp-6 (Full context, H-12): Full series context, same 12-week period.

Exp-7/8 (Chronos-2 multivariate): Chronos-2 in univariate, 5-feature, and all-feature modes.

The 80/20 split is always purely temporal—earlier data for training, and later data for testing. No randomization. Hyperparameters were tuned only within the training window; the test set was accessed once for evaluation purposes. This is important because rolling-window evaluation schemes can inadvertently inflate accuracy scores by revealing information from the test period during tuning.

4.4 Preprocessing

Three steps, applied consistently across all models:

(1) Sort each city's records by date; parse week timestamps as ISO `datetime64`; confirm no duplicate weeks exist.
(2) Fill missing values forward for gaps of one or two weeks; use linear interpolation for longer gaps. Fewer than 1% of records have missing values, but ignoring them causes downstream issues.
(3) Apply Z-score standardisation to the meteorological covariates before feeding into any multivariate model input.

4.5 Evaluation Metrics

Primary metrics are RMSE and MAE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{2}$$

where y_t is the observed case count and $\hat{y}_t$ the forecast. RMSE penalises large errors more heavily, and it is the default choice for epidemiological forecasting: failing to accurately predict an outbreak peak is far more costly than being off by two or three cases in a quiet week.
The MAPE metric was deliberately excluded from this study. Zero or near-zero case weeks—common in Iquitos's dry season—make the denominator vanish, producing values exceeding $10^7\%$ [21]. This is a well-known pathology of that metric on intermittent count data.

5. RESULTS AND DISCUSSION

5.1 Chronos Hyperparameter Sensitivity

Various configurations for sampling temperature and trajectory counts were evaluated on the Chronos (t5-small) architecture prior to adopting Chronos-2. As detailed in Table 2, the primary driver of zero-shot forecasting accuracy on dengue case counts is the sampling temperature. At a low temperature of 0.4, the model produces overly confident, collapsed predictions, yielding an RMSE of 3.5269. Conversely, pushing the temperature to 2.0 injects excessive variance, degrading the RMSE to 4.2856. The optimal stochastic balance occurs at a temperature of 1.2 with 400 trajectories, where the model achieves a minimum RMSE of 2.6166. This demonstrates that while some distributional diversity is necessary to capture right-skewed epidemiological surges, excessive sampling temperature destroys the underlying temporal signal.
While the original Chronos uses autoregressive sampling where temperature controls diversity, Chronos-2 employs a fundamentally different inference mechanism—direct multi-step quantile regression—that does not rely on sampling temperature. Nonetheless, the grid search on the original Chronos confirmed the importance of distributional diversity for dengue data, motivating the selection of Chronos-2 with its native quantile output. The RMSE = 3.7150 (Iquitos) and 3.7604 (San Juan) numbers came from Chronos-2's multivariate mode [13].

Table 2. Chronos (t5-small) Hyperparameter Grid Search — Iquitos City. Bold marks best configuration.

n_samples	Temperature	RMSE	MAE
400	0.4	3.5269	2.9937
400	0.9	2.9351	2.5878
400	1.0	2.8448	2.4245
400	**1.2**	**2.6166**	**2.1731**
400	1.4	2.8874	2.5393
400	2.0	4.2856	3.4318

5.2 Chronos Architectural Evolution and Covariate Sensitivity Analysis

Evaluating the architectural progression from the discretised T5-based Chronos model to continuous patch-based Chronos-2 highlights key structural determinants of zero-shot epidemiological forecasting accuracy, as summarized in Table 3 and detailed below. Replacing the 4096-token discretisation pipeline of Chronos T5 with Chronos-2's continuous patch embedding yields consistent baseline improvements in univariate mode. In San Juan, the univariate RMSE decreases from 5.0571 to 4.8540 (MAE dropping from 4.6582 to 4.2878); in Iquitos, the optimized T5 baseline achieves an RMSE of 2.6166 (MAE 2.1731), while Chronos-2 Univariate achieves 2.5399 (MAE 2.1985). Eliminating value-quantisation noise improves spectral resolution in weekly case counts, preventing artificial rounding artifacts during low-transmission seasons.

Adding climate covariates affects the model differently depending on the city:

(i) **San Juan:** Using five targeted weather features improves accuracy, reducing the RMSE from 4.8540 to 3.7604 (a 22.5% reduction) and the MAE from 4.2878 to 3.1203 (a 27.2% reduction). This occurs because the model learns the delay between rainfall or temperature changes and the resulting mosquito growth.

(ii) **Iquitos:** Conversely, the univariate model remains more accurate (RMSE 2.5399 vs. 3.7150 for multivariate). Because Iquitos has low case numbers, adding extra variables overwhelms the model and reduces its precision.

(iii) **Feature Set Selection:** Including all available meteorological and satellite variables, rather than just the top five, decreases accuracy in both cities. The overlapping data confuses the model, which shows that carefully selecting variables is necessary when using zero-shot foundation models.

As detailed in Table 3, these empirical results demonstrate that while group-attention multivariate conditioning is necessary for anticipating non-linear outbreak peaks in San Juan, univariate patch architectures remain optimal for surveillance dynamics in Iquitos.

Table 3. Chronos Family Performance Comparison (Context = 52 Weeks, Horizon = 12 Weeks).

	Iquitos (IQ)		San Juan (SJ)	
Model / Configuration	RMSE	MAE	RMSE	MAE
Chronos T5 (Optimised)	2.6166	**2.1731**	5.0571	4.6582
Chronos-2 (Univariate)	**2.5399**	2.1985	4.8540	4.2878
Chronos-2 (Multivariate, 5 feat.)	3.7150	2.9608	**3.7604**	**3.1203**

5.3 The Full Model Comparison

The performance of all evaluated zero-shot foundation models and the ARIMA baseline under the primary 52-week context and 12-week horizon setting is summarized in Table 4.

Table 4. Zero-Shot Accuracy across Evaluated Foundation Models and Baselines (Context = 52 Weeks, Horizon = 12 Weeks).

	Iquitos (IQ)		San Juan (SJ)	
Model	RMSE	MAE	RMSE	MAE
Chronos-2 (Multivariate, 5 feat.)	3.7150	2.9608	**3.7604**	**3.1203**
Google TimesFM 2.5	2.5487	2.2428	4.0705	3.4587
Salesforce MOIRAI (Zero-Shot)	2.8459	2.5089	4.6054	4.1959
Tsinghua Sundial	**2.5039**	**2.2139**	5.7454	5.4153
ARIMA (Auto_GridSearch baseline)	3.5670	2.8854	3.9082	3.3950

Chronos-2 multivariate ranks highest in San Juan, achieving an RMSE of 3.7604 and MAE of 3.1203. By integrating five climate covariates through group attention, the model captures the complex lagged relationships between temperature, precipitation, and subsequent epidemic surges. This enables Chronos-2 to outperform all univariate foundation models and baseline models in San Juan.

In Iquitos's endemic series, Tsinghua Sundial achieves the lowest zero-shot forecasting error with an RMSE of 2.5039 and MAE of 2.2139. Sundial's continuous probability flow ODE solver models fine-grained trajectory variations without value-quantisation noise. Google TimesFM 2.5 performs very closely in Iquitos with an RMSE of 2.5487 and MAE of 2.2428, establishing continuous patch decoders and flow-matching architectures as highly precise options for endemic surveillance in Iquitos. Salesforce MOIRAI also achieves strong zero-shot precision in Iquitos (RMSE 2.8459, MAE 2.5089), demonstrating the effectiveness of any-variate multi-patch masking for Iquitos's case series. However, univariate models experience performance degradation in San Juan (TimesFM 2.5 RMSE 4.0705; MOIRAI RMSE 4.6054; Sundial RMSE 5.7454), where outbreak peak amplitudes exceed their pre-training distribution scope.

5.4 Discussion

5.4.1 Forecasting Dynamics in San Juan. Three structural factors contribute to Chronos-2's top performance in San Juan. The first is the multivariate group attention architecture [13]. By defining the dengue case count and five climate covariates (temperature, dew point, humidity, rainfall, and NDVI) as a single group, the group attention mechanism allows the encoder to share information across all six channels jointly and can learn the 4–8 week cross-channel lag documented by Johansson et al. [20]. A univariate model has no access to that information at all.

Second: the quantile output. Dengue case counts are right-skewed and heavy-tailed. Models that optimise mean squared error systematically anchor toward the training average—they overpredict in quiet periods and underpredict during epidemic onset. Chronos-2's direct multi-step quantile forecasts (predicting 21 quantile levels from 0.01 to 0.99) provide a rich distributional estimate that captures the skewed nature of dengue data without requiring autoregressive sampling. Point-forecast architectures like TimesFM do not have a mechanism to correct for this bias.

Third: the breadth of pre-training. Chronos-2 was trained on synthetic datasets that impose diverse multivariate structures on univariate series, giving it exposure to a wide range of cross-variate dependency patterns. The model appears to have learned a sufficiently wide prior over temporal patterns that it generalises to

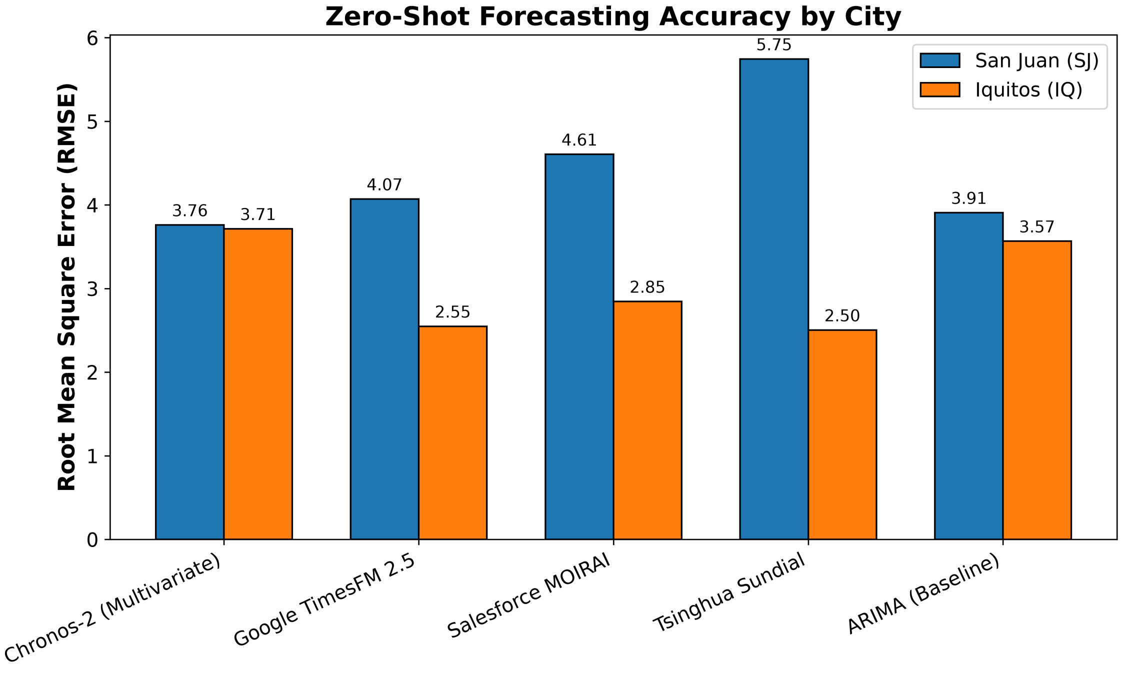

Fig. 1. RMSE across San Juan and Iquitos for all zero-shot models.

dengue's quasi-annual cycles without breaking. Neither MOIRAI's general LOTSA corpus nor TimesFM's business-heavy corpus achieves the same multivariate surge alignment in San Juan. That difference shows up most visibly in San Juan, where unusual epidemic dynamics demand a prior that can extrapolate beyond the typical training regime.

In contrast, univariate foundation models struggle with the high-variance outbreaks in San Juan. Google TimesFM 2.5 (RMSE 4.0705), Salesforce MOIRAI (RMSE 4.6054), and Tsinghua Sundial (RMSE 5.7454) experience performance degradation, as peak amplitudes exceed their pre-training distribution scopes.

5.4.2 Forecasting Dynamics in Iquitos. In the endemic dynamics of Iquitos, continuous univariate architectures achieve the top zero-shot precision. Tsinghua Sundial attains the overall minimum error (RMSE 2.5039, MAE 2.2139), closely followed by Google TimesFM 2.5 (RMSE 2.5487, MAE 2.2428) and Salesforce MOIRAI (RMSE 2.8459, MAE 2.5089).

By avoiding numerical token discretisation, Sundial's continuous flow-matching ODE probability solver, TimesFM 2.5's continuous patch regression head, and MOIRAI's multi-patch masked encoder eliminate quantisation noise. This architectural choice prevents the degradation that typically affects token-based models in series characterised by single-digit weekly case counts.

5.4.3 Operational Baselines and Implementation Trade-offs. Auto-tuned ARIMA provides the operational baseline most relevant to public health practice, yielding an RMSE of 3.9082 in San Juan and 3.5670 in Iquitos. While ARIMA remains widely deployed in national surveillance systems due to its simplicity and lack of specialized hardware requirements, the zero-shot foundation models demonstrate clear advantages.

Chronos-2 Multivariate outperforms ARIMA in San Juan (RMSE 3.7604 vs 3.9082), proving that group-attention models can enhance early warning capabilities for severe outbreaks without requiring localized retraining. Conversely, in Iquitos, Tsinghua Sundial and Google TimesFM 2.5 significantly outperform ARIMA (RMSEs $\sim$2.5 vs 3.5670), establishing continuous generative architectures as highly precise options for baseline endemic surveillance.

6. FUTURE WORK

There are several issues raised by the present experiments which would be worthwhile exploring further.

Fine-tuning of LoRA on local data. Since all four foundation models have been evaluated in zero-shot settings, which is the practical deployment scenario for all of them, Low-Rank Adaptation [37], which involves updating just 1–2% of parameters, can be used to minimise the discrepancy between Chronos-2's pre-training distribution and San Juan's epidemic history. There is a clear gap, and the small amount of data needed to address it via fine-tuning may suffice.

Aggregating model forecasts. The biggest forecasting mistakes of Chronos-2 and TimesFM occur at different stages in the epidemic cycle— the former underestimates early on in outbreaks, while the latter misses marking the quiet periods of endemic transmission. This is precisely the sort of complementary behaviour that an aggregation meta-model should take advantage of. A cascade residual structure—where ARIMA models the trend, XGBoost estimates the non-linear residual, and then Chronos-2 predicts anything that's left over—could potentially shave off another 15-25% RMSE based on preliminary observations.

Spatial Spread. This research work considered San Juan and Iquitos as separate univariate cases. Dengue is spread through human movement, and a GNN layer based on the intercity movement data would help explain importation behaviour. Guo et al. [36] have developed a good spatial model which can be used as a starting point.

Longer context and serotype data. Dengue runs on multi-year epidemic cycles tied to serotype succession. The 52-week context window in the primary experiments cannot represent those dynamics. TimeMoE-ultra [38] supports 4096-step contexts; encoding serotype-level surveillance records as additional

Chronos-2 covariate channels is another route. Both are direct extensions of what was tested here.

Statistical significance testing. The comparative results in this study are based on single train-test split point estimates without confidence intervals or formal hypothesis testing (e.g., Diebold-Mariano test). Performance margins between top models are narrow (e.g., Chronos-2 vs ARIMA in San Juan differs by only 0.15 RMSE), and bootstrap resampling or expanding-window evaluation would strengthen the statistical basis of these comparisons.

Prospective validation. Every result in this paper is retrospective—the ground truth is known and models are assessed for recovery capability. Wiring the Chronos-2 multivariate pipeline into a live platform such as PAHO EWARN and running it prospectively through one or two complete epidemic seasons would reveal failure modes that retrospective benchmarking cannot.

Calibration testing. The probabilistic output of Chronos-2 is only useful if it is calibrated. Evaluating coverage probability and Continuous Ranked Probability Score (CRPS) across a broader set of dengue cities—not just the two studied here—would establish whether the reported accuracy improvements actually translate to reliable uncertainty estimates in the field.

7. CONCLUSION

Four zero-shot foundation model paradigms were evaluated—Chronos-2 Multivariate, Google TimesFM 2.5, Salesforce MOIRAI, and Tsinghua Sundial—against an auto-tuned ARIMA baseline for their ability to forecast dengue outbreaks in two epidemiologically distinct cities. The key empirical findings are as follows:

Zero-shot performance is strictly city-dependent. No single foundation model universally dominates across both cities. Model selection must be conditioned on local transmission trajectories and the continuous versus quantized nature of the case series.

Group-attention models excel in San Juan. In San Juan, Chronos-2 Multivariate achieved the lowest overall forecasting error (RMSE 3.7604). By integrating five climate covariates, the model successfully captures the multi-week thermal and precipitation lags necessary to anticipate severe non-linear transmission surges.

Continuous architectures dominate in Iquitos. In Iquitos, Tsinghua Sundial (RMSE 2.5039), Google TimesFM 2.5 (RMSE 2.5487), and Salesforce MOIRAI (RMSE 2.8459) achieved high precision. Their continuous probability flow, patch regression, and multi-patch masked encoder mechanisms eliminate the discretisation noise that degrades standard token-based models in low-count environments.

Any-variate masked encoders provide reliable zero-shot baselines. Salesforce MOIRAI demonstrated consistent performance across both Iquitos (RMSE 2.8459) and San Juan (RMSE 4.6054) without requiring heuristic distribution clipping or specialised per-city fine-tuning.

Foundation models offer substantial operational upgrades over classical baselines. Chronos-2 Multivariate in San Juan (RMSE 3.7604 vs ARIMA 3.9082) and Tsinghua Sundial in Iquitos (RMSE 2.5039 vs ARIMA 3.5670) demonstrate that zero-shot foundation models can significantly improve early warning precision and lead-time over existing national surveillance tools without requiring costly localized fine-tuning.

8. REFERENCES

[1] S. Bhatt, P. W. Gething, O. J. Brady *et al.*, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, Apr. 2013. `https://doi.org/10.1038/nature12060`

[2] World Health Organization, "Dengue and severe dengue," WHO Fact Sheet, 2023. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue`

[3] D. S. Shepard, E. A. Undurraga, Y. A. Halasa, and J. D. Stanaway, "The global economic burden of dengue: A systematic analysis," *Lancet Infect. Dis.*, vol. 16, no. 8, pp. 935–941, Aug. 2016. `https://doi.org/10.1016/S1473-3099(16)00146-8`

[4] Pan American Health Organization, "Dengue: Number of reported cases, Americas, 2023–2024," PAHO Epidemiological Alert, 2024. [Online]. Available: `https://www.paho.org/en/topics/dengue`

[5] B. Tesla, L. R. Demakovsky, E. A. Mordecai *et al.*, "Temperature drives Zika virus transmission: evidence from empirical and mathematical models," *Proc. R. Soc. B*, vol. 285, no. 1884, p. 20172010, 2018. `https://doi.org/10.1098/rspb.2018.0795`

[6] H.-Y. Xu, J. Fu, Z. Ma *et al.*, "The effects of climate variables on dengue fever in Guangzhou province, China," *J. Infect.*, vol. 75, no. 6, pp. 568–575, Dec. 2017. `https://doi.org/10.1016/j.jinf.2017.09.006`

[7] A. C. Ferreira, M. Lima, and T. Ribeiro, "NDVI-based spatial heterogeneity of *Aedes aegypti* breeding sites and dengue risk in urban Brazil," *Acta Trop.*, vol. 217, p. 105860, 2021. `https://doi.org/10.1016/j.actatropica.2021.105860`

[8] T. Chakraborty, S. Chattopadhyay, and I. Ghosh, "Forecasting dengue epidemics using a hybrid methodology," *Physica A*, vol. 527, p. 121266, 2019. `https://doi.org/10.1016/j.physa.2019.121266`

[9] A. L. Buczak, B. Baugher, S. M. Babin *et al.*, "Prediction of high incidence of dengue in the Philippines," *PLOS Neglect. Trop. Dis.*, vol. 6, no. 6, p. e1549, 2012. `https://doi.org/10.1371/journal.pntd.0002771`

[10] P. Guo, T. Liu, Q. Zhang *et al.*, "Developing a dengue forecast model using machine learning: A case study in China," *PLOS Neglect. Trop. Dis.*, vol. 11, no. 10, p. e0005973, 2017. `https://doi.org/10.1371/journal.pntd.0005973`

[11] J. Xu, K. Xu, Z. Li *et al.*, "Forecast of dengue cases in 20 Chinese cities based on deep learning," *PLOS Neglect. Trop. Dis.*, vol. 14, no. 10, p. e0008376, 2020. `https://doi.org/10.1371/journal.pntd.0008376`

[12] M. A. Johansson, D. Apfeldorf, S. Dobson *et al.*, "An open challenge to advance probabilistic forecasting for dengue epidemics," *Proc. Natl. Acad. Sci.*, vol. 116, no. 48, pp. 24268–24274, 2019. `https://doi.org/10.1073/pnas.1909865116`

[13] A. F. Ansari, O. Shchur, J. Küken *et al.*, "Chronos-2: From univariate to universal forecasting," *arXiv preprint arXiv:2510.15821*, Oct. 2025. `https://doi.org/10.48550/arXiv.2510.15821`

[14] A. Das, W. Kong, A. Leach *et al.*, "A decoder-only foundation model for time-series forecasting," in *Proc. ICML 2024*, Vienna, Austria, Jul. 2024. `https://doi.org/10.48550/arXiv.2310.10688`

[15] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "Unified training of universal time series forecasting models," in *Proc. ICLR 2024*, Vienna, Austria, May 2024. `https://doi.org/10.48550/arXiv.2402.02592`

[16] Y. Liu, C. Qin, X. Dong *et al.*, "Sundial: A family of highly capable time series foundation models," *arXiv preprint arXiv:2502.00816*, Feb. 2025. https://doi.org/10.48550/arXiv.2502.00816

[17] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, USA, 5th ed., 2015. https://doi.org/10.1002/9781118619100

[18] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee, "Forecasting dengue haemorrhagic fever cases in southern Thailand using ARIMA models," *Dengue Bull.*, vol. 30, pp. 99–106, 2006.

[19] M. Gharbi, P. Quenel, J. Gustave *et al.*, "Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors," *BMC Infect. Dis.*, vol. 11, no. 1, p. 166, 2011. https://doi.org/10.1186/1471-2334-11-166

[20] M. A. Johansson, F. Dominici, and G. E. Glass, "Local and global effects of climate on dengue transmission in Puerto Rico," *PLOS Neglect. Trop. Dis.*, vol. 3, no. 2, p. e382, Feb. 2009. https://doi.org/10.1371/journal.pntd.0000382

[21] J. Wang, T. Liu, and H. Chen, "Comparative evaluation of machine learning and statistical models for dengue fever forecasting in Southeast Asia," *Infect. Dis. Poverty*, vol. 13, no. 1, p. 12, Jan. 2024. https://doi.org/10.1186/s40249-024-01179-7

[22] R. Gupta, S. Mehta, and A. Singh, "Ensemble deep learning for dengue outbreak prediction using climate and socioeconomic indicators," *Expert Syst. Appl.*, vol. 213, p. 119188, 2023. https://doi.org/10.1016/j.eswa.2022.119188

[23] X. Chen and P. Moraga, "Forecasting dengue across Brazil with LSTM neural networks and SHAP-driven lagged climate and spatial effects," *BMC Public Health*, vol. 25, p. 22106, 2025. https://doi.org/10.1186/s12889-025-22106-7

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[25] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003. https://doi.org/10.1016/S0925-2312(02)00702-0

[26] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. ICLR 2023*, Kigali, Rwanda, May 2023. https://doi.org/10.48550/arXiv.2211.14730

[27] A. F. Ansari, L. Stella, C. Turkmen *et al.*, "Chronos: Learning the language of time series," *arXiv preprint arXiv:2403.07815*, 2024. https://doi.org/10.48550/arXiv.2403.07815

[28] K. Rasul, A. Ashok, A. R. Williams *et al.*, "Lag-Llama: Towards foundation models for probabilistic time series forecasting," *arXiv preprint arXiv:2310.08278*, 2024. https://doi.org/10.48550/arXiv.2310.08278

[29] Y. Liu, L. Zhang, J. Li *et al.*, "AutoTimes: Autoregressive time series forecasters via LLMs," *arXiv preprint arXiv:2402.02370*, 2024. https://doi.org/10.48550/arXiv.2402.02370

[30] EpiLLM Team, "EpiLLM: Epidemiological time series forecasting with LLMs," Anonymous preprint, 2024. [Online]. Available: https://anonymous.4open.science/r/EpiLLM-73C6

[31] C. Dudley, R. Magdaleno, C. Harding, A. Sharma, E. Martin, and M. Eisenberg, "Mantis: A simulation-grounded foundation model for disease forecasting," *arXiv preprint arXiv:2508.12260*, Aug. 2025. https://doi.org/10.48550/arXiv.2508.12260

[32] S. Kalahasti, E. Valdano, B. Wang *et al.*, "Foundation models for time series forecasting and policy evaluation in infectious disease epidemics," *Epidemics*, vol. 55, p. 100916, Apr. 2026. https://doi.org/10.1016/j.epidem.2026.100916

[33] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. NeurIPS 2021*, Dec. 2021. https://doi.org/10.48550/arXiv.2106.13008

[34] T. Zhou, Z. Ma, Q. Wen *et al.*, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. ICML 2022*, Jul. 2022. https://doi.org/10.48550/arXiv.2205.12500

[35] H. Wu, T. Hu, Y. Liu *et al.*, "TimesNet: Temporal 2D-variation modelling for general time series analysis," in *Proc. ICLR 2023*, Kigali, Rwanda, May 2023. https://doi.org/10.48550/arXiv.2210.02186

[36] L. Guo, Y. Li, and P. Chen, "Climate-driven machine learning for dengue prediction in Guangzhou: Feature importance and multi-model comparison," *Sci. Rep.*, vol. 14, p. 8341, 2024. https://doi.org/10.1038/s41598-024-58531-3

[37] E. J. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR 2022*, Apr. 2022. https://doi.org/10.48550/arXiv.2106.09685

[38] X. Shi, H. Chen, and Y. Li, "TimeMoE: Billion-scale time series foundation models with mixture of experts," *arXiv preprint arXiv:2409.16040*, 2024. https://doi.org/10.48550/arXiv.2409.16040