A High-Resolution (1-km) Ecological Niche Model for Dengue Vectors in South Asia

Radhana Kriti Mainali
Poolesville High School, Poolesville, Maryland

ABSTRACT

Dengue fever is mainly spread by *Aedes aegypti* and *Aedes albopictus*, two mosquito species that have been moving rapidly into new tropical and subtropical regions. Because of this expansion, having reliable maps of where these vectors can survive is important for estimating disease risk. In this project, I built a preliminary ecological niche model by combining occurrences of both dengue vectors in South Asia using MaxEnt.

I compiled 383 occurrences for *Ae. aegypti* and 129 for *Ae. albopictus* from published datasets and paired them with 48 environmental predictors, including bioclimatic variables from WorldClim, climate indices from ENVIREM, population density from WorldPop, and land-cover information from EarthEnv. All layers were available at a 30-arc-second (~1 km) resolution. After filtering for multicollinearity, 12 predictors remained in the final model set. I trained the MaxEnt model (evaluated across 125 candidate runs) using 5-fold cross-validation with 10,000 randomly generated background points, and I converted predictions to binary outputs using a 0.5 threshold. The mean AUC across candidate models was 0.857.

The highest predicted suitability occurred in low-lying, densely populated, and coastal regions, and moderate suitability was found in more transitional environments such as the Himalayan foothills. Variable-importance analyses showed three dominant predictors: mean temperature of the driest quarter, isothermality, and precipitation of the driest month. This preliminary study still has several limitations (to be resolved in the coming months), including the fact that all occurrence data came only from South Asia, that the environmental layers approximate microhabitat conditions fairly roughly, and that socioeconomic factors were missing. In the future, I plan to expand the occurrence dataset to a global scale, incorporate finer-resolution predictors, and experiment with newer approaches such as AlphaEarth foundation-model embeddings and deep-learning architectures. These steps should help increase spatial accuracy and provide more detailed ecological insight to support dengue vector monitoring and public-health planning.

INTRODUCTION

Dengue fever is mainly transmitted by two mosquito vectors: *Aedes aegypti* and *Aedes albopictus*. Over the past 50 years, both have vastly expanded their geographic ranges (Kraemer et al., 2015), making dengue an increasingly important public health concern in many tropical and even subtropical regions. *Ae. aegypti* thrives in close association with humans and is most common in urban and peri-urban areas, while *Ae. albopictus* can tolerate a wider range of environments, including suburban and rural areas (Messina et al., 2015). Because of these expanding ranges and their differing habitat preferences, mapping where each species can survive is essential for understanding current dengue risk and planning interventions (Bhatt et al., 2013; Kraemer et al., 2015).

Several global mapping studies have used machine learning to predict vector presence and dengue risk (Bhatt et al., 2013; Messina et al., 2015). However, these often rely on environmental predictors at coarse spatial resolutions (around 25 km^2), which can blur important local differences in climate, land cover, and human activity. In this project, I use MaxEnt, a presence-only modeling method, at a 1 km^2 spatial resolution to capture the small-scale environmental patterns that may influence the distributions of *Ae. aegypti* and *Ae. albopictus* in South Asia. This is a preliminary analysis limited to regional occurrence data. I plan to add global records in the next stage to capture a broader range of environmental conditions and improve niche characterization. I also expect to involve additional collaborators as the project grows and the modeling becomes more detailed.

METHODS

Occurrence Data

Occurrences were collected from Kraemer et al. (2015) and clipped to South Asia (India, Pakistan, Bangladesh, Nepal, Sri Lanka). Only presence records were used.

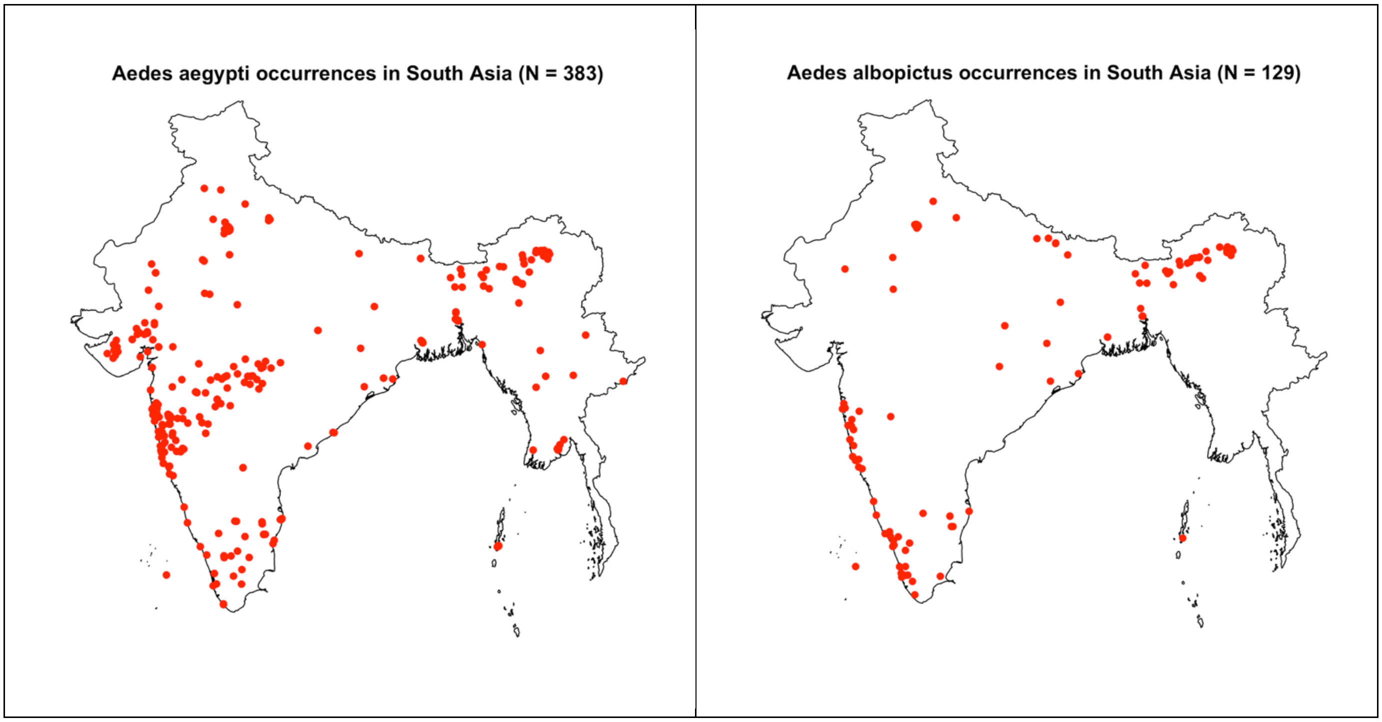
Aedes aegypti occurrences in South Asia (N = 383)
Aedes albopictus occurrences in South Asia (N = 129)

Figure 1. Distributional records of two vectors of dengue in South Asia.

Environmental Predictors

All environmental layers were processed at a spatial resolution of 30 arc-seconds (~1 km^2 at the equator). The dataset included 48 covariates in total: 19 bioclimatic variables from WorldClim, 16 ENVIREM-derived climatic indices, 1 population density layer, and 12 land cover layers. The covariates spanned four thematic categories:

- **Bioclimatic Variables (WorldClim v2.1):** 19 standard predictors used in ecological niche modeling representing annual trends, seasonality, and limiting environmental factors (bio1–bio19). The WorldClim version 2.1 climate data for 1970-2000 was downloaded from https://www.worldclim.org/data/worldclim21.html.
- **Derived Climate Indices (ENVIREM):** 16 variables of PET and moisture indices relevant for vector ecology (e.g., annualPET, aridityIndexThornthwaite, embergerQ). The ENVIREM (Environmental Rasters for Ecological Modeling) dataset was downloaded from https://envirem.github.io.
- **Population Density:** Derived from WorldPop adjusted to UN national totals, ~1 km resolution (WorldPop, 2016). The population data came from WorldPop 2016, and was downloaded from https://hub.worldpop.org/doi/10.5258/SOTON/WP00013.
- **Land Cover:** The land-cover predictor was obtained from the EarthEnv Global 1-km Consensus Land Cover dataset, which integrates multiple remote-sensing land-cover products into a unified 1-km global layer (Tuanmu & Jetz, 2014). The dataset published at https://www.earthenv.org/landcover was downloaded with Google Earth Engine API.

Temperature and precipitation related predictors were included because mosquito development and survival are climate-sensitive. Population density was included because human hosts strongly influence mosquito presence. Land-cover variables were included because vegetation, built-up areas, and water-related features affect mosquito breeding habitat and micro-environmental conditions.

Raster layers were resampled, aligned, and masked to the extent of the study area. Variables with Pearson's r > 0.75 were filtered to reduce collinearity. This retained only 12 out of 48 initial covariates: bio 13, bio 14, bio 15, bio 18, bio 19, bio 3, bio 9, aridityIndexThornthwaite, PETseasonality, PETWettestQuarter, pop_mean, cultivated (**Table 1**).

Table 1. List of 12 retained covariates out of the original 48 after multicollinearity filtering with the threshold of Pearson's correlation of absolute value of 0.75.

SN	Covariate Name	Covariate Full Name
1	bio_13	Precipitation of Wettest Month
2	bio_14	Precipitation of Driest Month
3	bio_15	Precipitation Seasonality (Coefficient of Variation)
4	bio_18	Precipitation of Warmest Quarter
5	bio_19	Precipitation of Coldest Quarter
6	bio_3	Isothermality
7	bio_9	Mean Temperature of Driest Quarter
8	aridityIndexThornthwaite	Thornthwaite Aridity Index
9	PETseasonality	Potential Evapotranspiration (PET) Seasonality
10	PETWettestQuarter	PET of Wettest Quarter
11	pop_mean	Mean Human Population Density
12	cultivated	Proportion of Cultivated Land Cover

Modeling

In this initial stage, occurrences of both Aedes aegypti and Aedes albopictus were pooled and modeled jointly to generate a combined suitability surface. This approach treats both vectors as a single presence dataset, allowing the model to capture broad environmental determinants shared between them. Four algorithms were initially considered for species distribution modeling: **Generalized Additive Models (GAM), Random Forest (RF), Gradient Boosting Machines (GBM), and Maximum Entropy (MaxEnt**). Although results were generated from all four, this report focuses exclusively on MaxEnt due to its alignment with the presence-only nature of the available data and its robust performance with small sample sizes. In contrast, the other algorithms conceptually require presence–absence or presence–pseudoabsence data structures, which were not appropriate for this dataset.

MaxEnt was implemented using the **ENMeval package in R**. For each run, 10,000 background points were randomly sampled across the study region. The modeling workflow included two levels of replication: five different random splits of background points and five split of five-fold cross-validation of presence data (125 models in total). This dual replication design allowed for robust evaluation of model performance across different environmental and cross-validation partitions. Feature classes evaluated included Linear and Quadratic (LQ), and regularization multipliers (RM) ranged from 0.5 to 4.0. The model with the lowest Akaike Information Criterion corrected for small sample sizes (AICc) was selected for final analysis.

The model was evaluated using 5-fold cross-validation, and predictions were converted to presence/absence using a threshold of 0.5. The cross-validated AUC values were generally high, indicating that the model performed well in distinguishing suitable from unsuitable areas.

Sampling Bias

The occurrence data are presence-only and may be biased toward populated areas, so the algorithm used only one record per grid cell to reduce clustering, and background points were generated randomly across the study region.

Evaluation

AUC and AICc were used to evaluate model fit. Final ensemble maps were generated by averaging predictions across the best model replicates in each combination of background draw and split (there are 25 such best models).

RESULTS

Model Selection

All best-performing models used LQ features and RM = 0.5. Feature class frequency and regularization multiplier frequency plots confirm this consistency. Across all 125 models, no other feature class or RM setting yielded lower AICc or more stable AUC performance, indicating that this simple model structure was well-suited to the data and avoided overfitting. This also highlights that more complex feature combinations (e.g., Hinge or Product) did not improve model performance for the presence-only dataset under current standard cross-validation and replication design.

Predictive Performance

Mean AUC was 0.857 for the best models of the combination of background sampling (5) and split (5) (**Figure 2**). The small standard deviation highlights the stability of the models to background variability and random cross-validation partitions. Additionally, low AICc scores were aligned with the highest AUC scores, further confirming the model selection criteria.

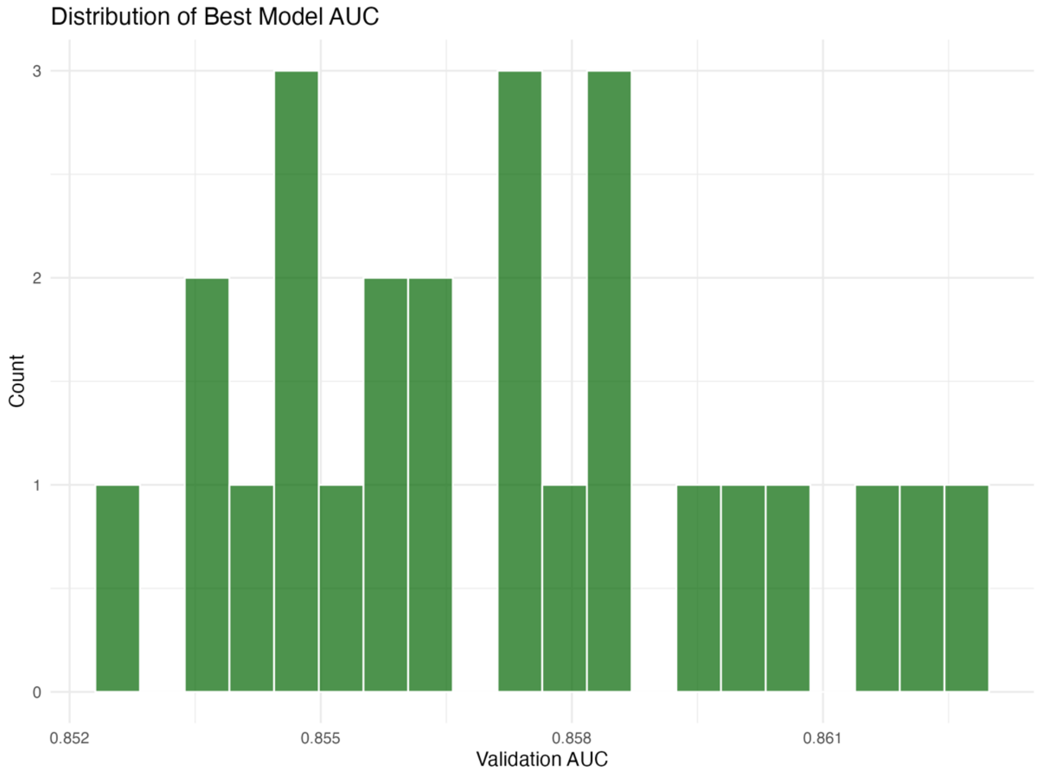

Figure 2. Histogram of AUC scores from the best MaxEnt models from each cross-validation partition (25 models in total).

Spatial Predictions

Figure 3 displays the predicted mean habitat suitability for the combined vector dataset. The highest suitability is observed in major urbanized and low-lying coastal areas (e.g., Bangladesh delta, western coast of India), as well as broader and more continuous/isolated predicted suitability, extending well into central India, and hilly terrain of Nepal. Binary classification maps (**Figure 4**) demonstrate the conservative delineation of suitable habitat.

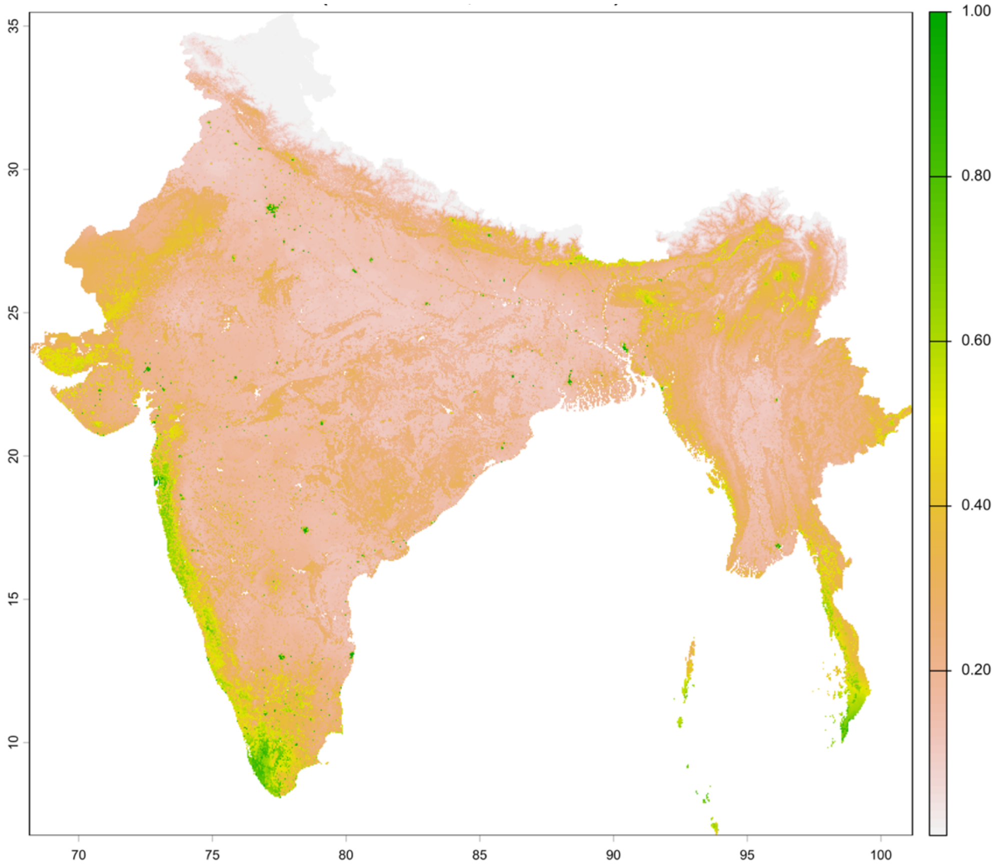

Figure 3. Ensemble mean prediction map of habitat suitability for *both vectors* using environmental layers at 30-arc-second resolution (≈1 km at the equator). All predictors and model outputs use the same resolution and were mapped in a geographic coordinate system (WGS84). No resampling beyond this native resolution was applied. Warmer tones indicate higher suitability.

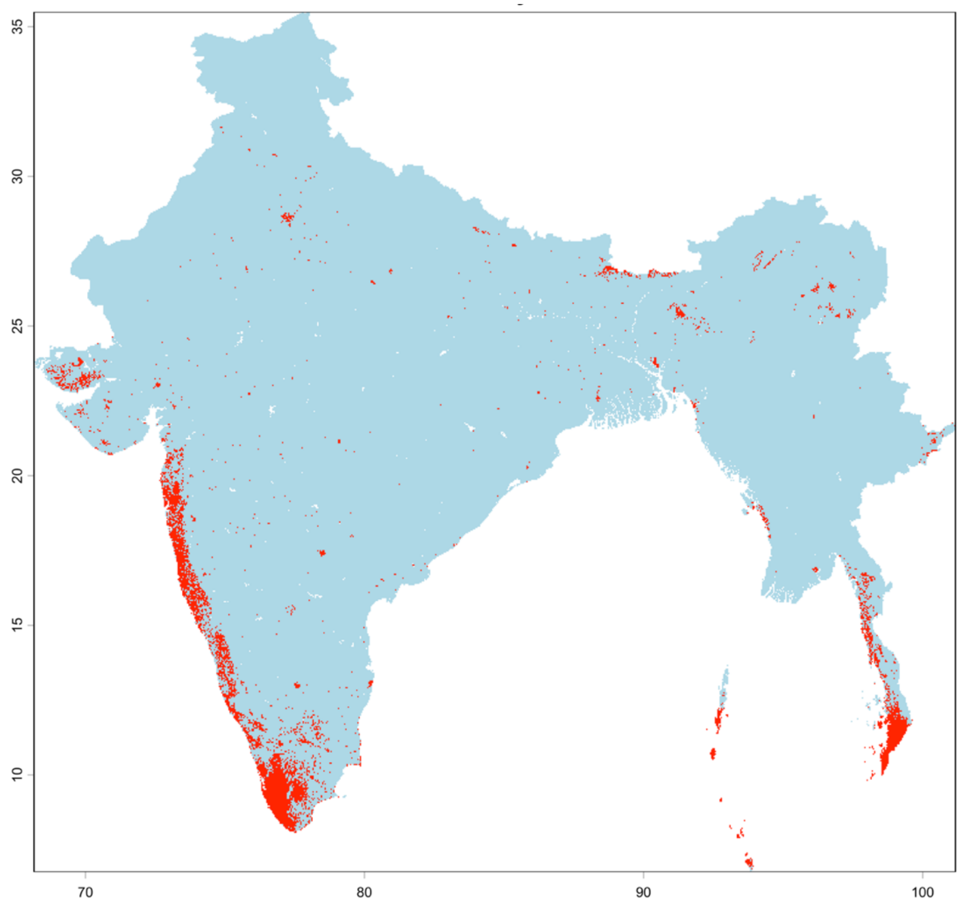

Figure 4. Binary presence–absence maps created using the 0.5 suitability threshold. These maps refine predictions to identify areas exceeding a conservative suitability cutoff.

Coefficient of variation (CV) maps (**Figure 5**) quantify spatial uncertainty across replicates. High-confidence zones (low CV) overlap strongly with major hotspots, affirming model reliability.

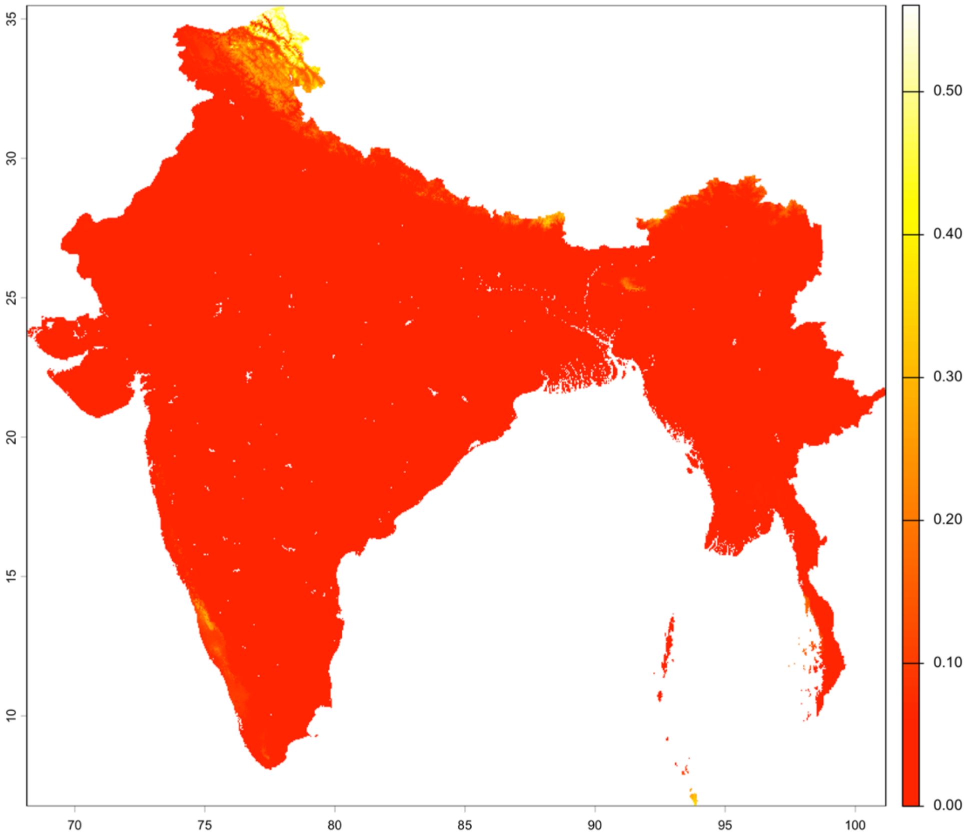

Figure 5. Coefficient of variation across 25 model predictions (best of each cross validation chosen). Darker shades represent higher model disagreement, most commonly in fringe areas with marginal suitability.

Variable Importance

Two complementary approaches were used to assess variable importance across the top 25 performing MaxEnt models. The first method calculates the mean and standard deviation of permutation importance across models (**Figure 6**). This approach can capture the influence of a variable through both its linear form and its squared (non-linear) transformations when selected by the model. For example, a predictor such as bio_14 may appear as both a linear term (bio_14) and a squared term (I(bio_14^2)), indicating that its influence may be non-linear. The second method (**Figure 7**) computes the total importance of each variable based on the sum of absolute coefficients from fitted model weights. In this case, the contributions of both raw and transformed (e.g., squared) terms of a given variable are aggregated to reflect its overall impact on model predictions. Both approaches consistently identified bio_9 (Mean Temperature of Driest Quarter), bio_3 (Isothermality), and bio_14 (Precipitation of Driest Month) as key predictors.

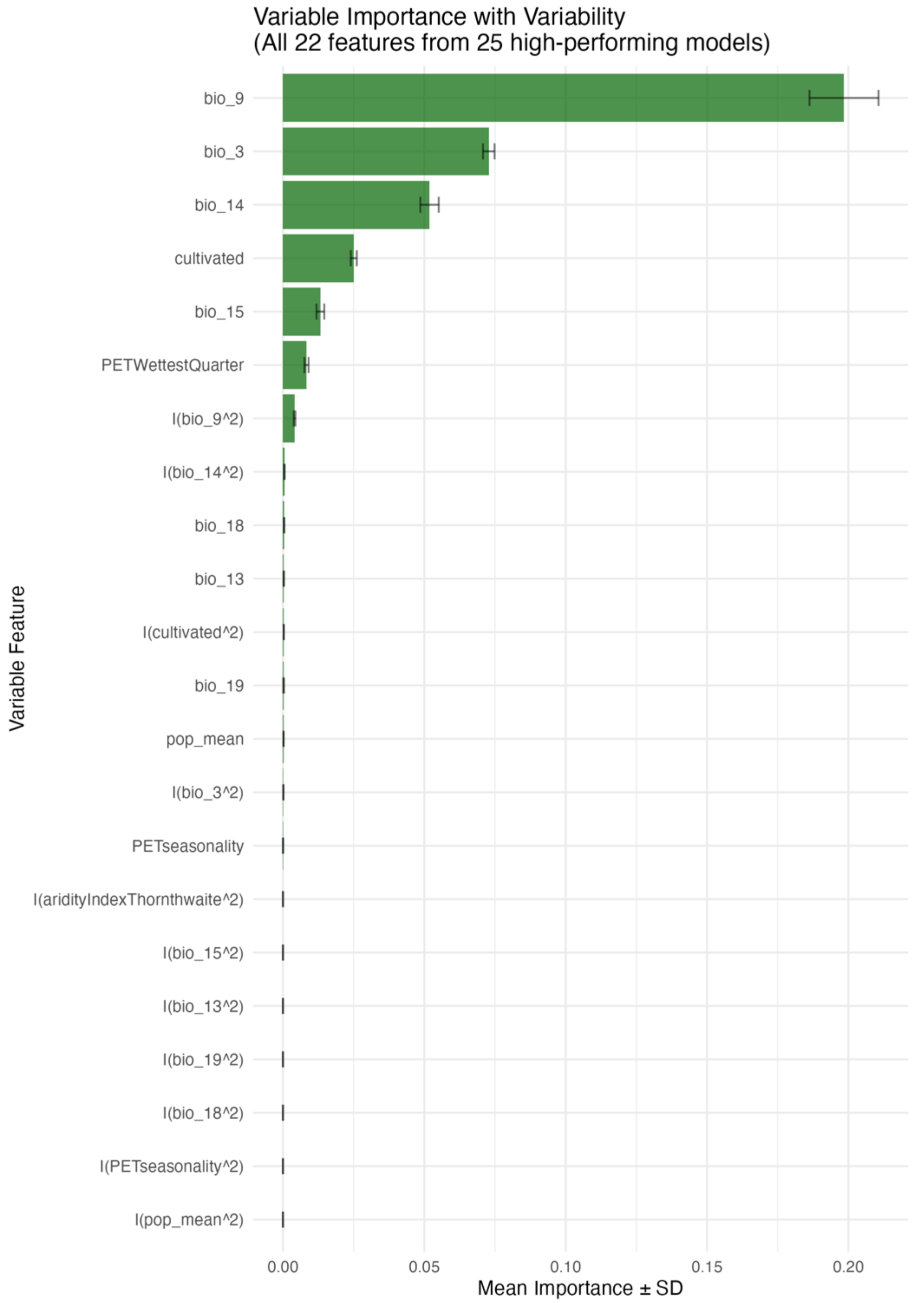

Figure 6. Mean permutation importance and variability of features across 25 models. Top predictors include temperature and moisture seasonality metrics. Bars represent mean ± SD.

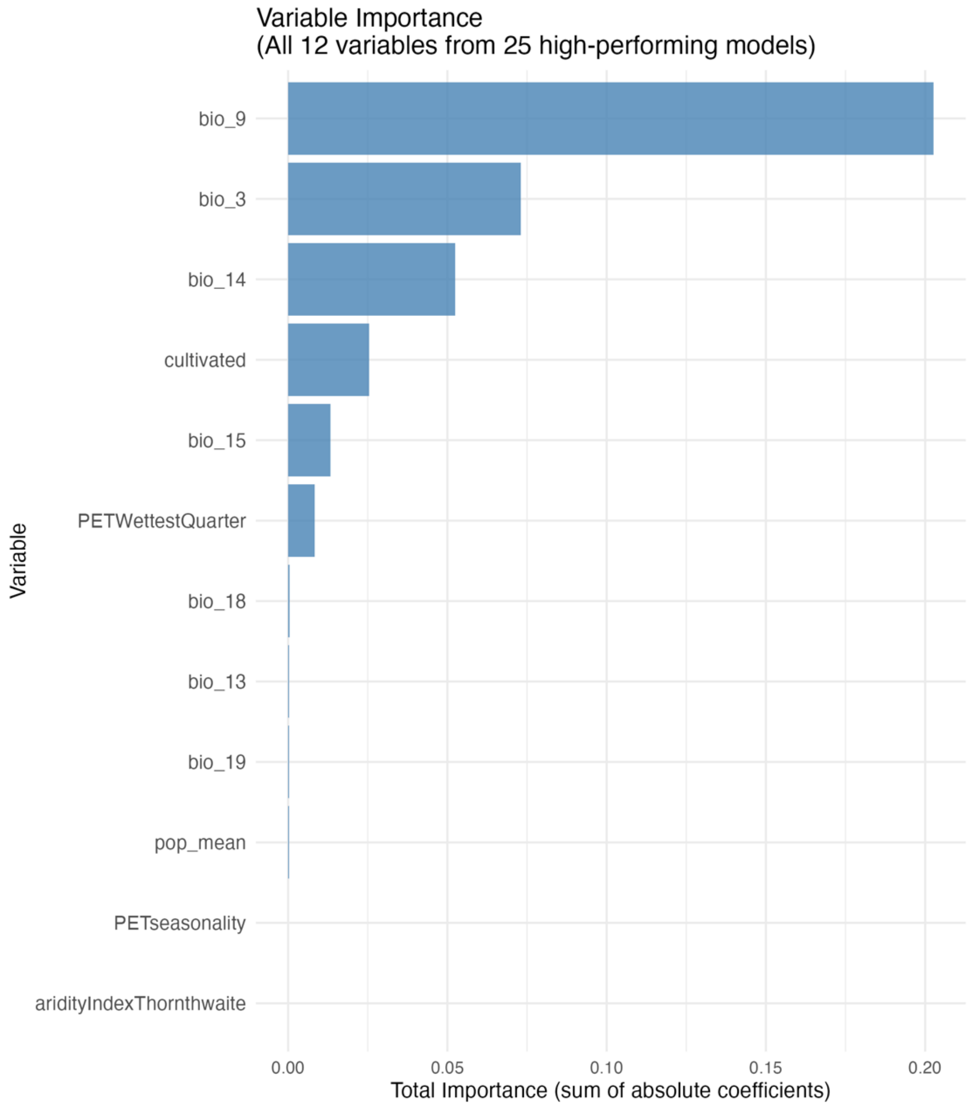

Figure 7. Total variable importance based on model coefficients. This emphasizes variables contributing to predictive structure across multiple models.

Response Curves for Key Predictors

Figure 8 displays the response curves for the six most influential predictors identified from the ensemble MaxEnt models. These plots illustrate how predicted suitability for vector presence varies as a function of individual environmental covariates, holding all other variables constant. Interestingly, diverse response shapes, indicating both linear and non-linear ecological effects were found. For instance, suitability exhibits a unimodal response with cultivated land proportion but bio_15 (precipitation seasonality) shows a steady positive association. In contrast, predictors like bio_9 (mean temperature of driest quarter) and bio_14 (precipitation of driest month) exhibit pronounced unimodal relationships, revealing thermal and moisture optima for vector persistence.

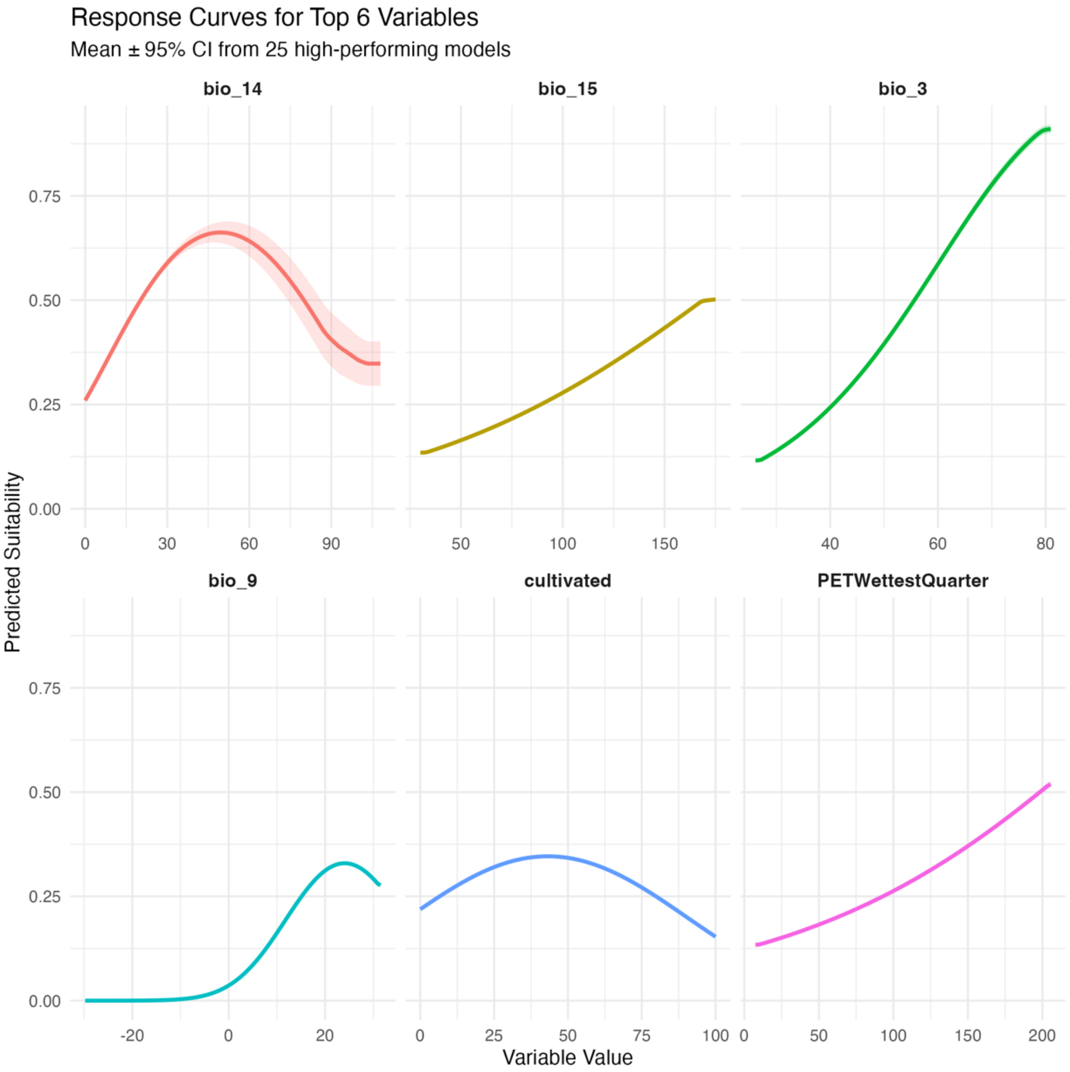

Figure 8. Response curves for the top 6 predictors derived from ensemble MaxEnt models. Solid lines represent the mean prediction and shaded bands reflect ±95% confidence intervals across 25 model replicates. Units are scaled to each variable's range.

DISCUSSION

Ecological Insight from Spatial Predictions

Predicted distributions reflect well-documented habitat preferences: strong urban affinity, vegetated and peri-urban landscapes, etc. The transition zones (e.g., the Himalayan foothills and the Western Ghats) exhibited moderate suitability with higher model disagreement (**Figure 5**), indicating these regions may serve as ecological boundaries or fronts for vector expansion.

Importance of Resolution

A key innovation of this study lies in its 1 km^2 spatial resolution. In contrast, most previous dengue or vector mapping studies, including Bhatt et al. (2013) and Kraemer et al. (2015), used coarser grids with spatial units of 25 km^2 or larger. These coarse aggregations dilute local heterogeneity, obscuring microhabitat cues essential for vector breeding. Our finer resolution enables detection of sub-kilometer habitat variation, distinguishing risk across neighboring urban blocks. This is because the finer resolution captures intra-urban and sub-urban heterogeneity and small-scale water-rich pockets missed in coarser resolutions. These finer patterns are especially relevant in densely populated areas where vector breeding often hinges on hyper-local infrastructure or household-level practices.

Microhabitat and Environmental Drivers

Vector suitability is determined by household-level conditions: uncovered water containers, domestic gardens, drainage ditches, and shaded refuse sites. These features are invisible in coarse environmental rasters but influential at 1 km^2 or finer scales. Although current land cover and population rasters act as proxies, future efforts should incorporate impervious surface area, road density, and waste disposal patterns derived from remote sensing or local data.

Limitations and Future Directions

One clear limitation of this early stage of the project is that I only used occurrence data from South Asia. That means the model is learning the ecological niche based on a relatively narrow range of climates and environments, which probably misses parts of the broader ecological tolerances of both species. The small sample size outside major urban areas also makes it harder to capture the full range of conditions these mosquitoes can handle. As I expand the analysis, I plan to bring in global records so the models can learn from a wider variety of temperature and precipitation settings and reduce regional bias.

At 1 km^2 resolution, models begin to pick up microhabitat structure, including peridomestic vegetation, small water bodies, slums with open containers, and shaded gardens (Messina et al., 2014; Bhatt et al., 2013). Coarse-resolution smoothing hides these details, introducing statistical artifacts. In addition, the model relied only on environmental variables; it did not include socioeconomic factors, human mobility, breeding-site density, or infrastructure data that influence urban mosquito populations. Future work will incorporate finer-resolution environmental predictors, including the Google AlphaEarth foundation model embeddings, which encode rich multispectral and temporal features useful for ecological modeling

(Google DeepMind, Google Research, & collaborators, 2025). These high-capacity predictors may allow the model to capture micro-environmental variation that is invisible at 1-km resolution.

I also plan to test deep-learning–based architectures that can learn nonlinear spatial patterns directly from high-resolution inputs, providing an alternative to traditional feature-based SDMs. In future, I plan to use a systematic spatial CV to reduce bias, and more robust thresholding methods, such as maximizing sensitivity and specificity, to reduce overprediction and improve classification reliability. I plan to build separate ecological niche models for *Ae. aegypti* and *Ae. albopictus*. This will help characterize species-specific environmental responses and compare their ecological niches more precisely.

This work is ongoing. As the project grows, I will involve additional collaborators to help extend the modeling to larger regions and refine the ecological interpretation.

Conclusion

This preliminary model demonstrates the value of high-resolution ecological niche modeling for dengue vectors. The results are still very much a first step, and there is a lot of room to improve the model reliability by adding global occurrences and predictors that capture finer-scale urban conditions. I will continue developing and expanding the project, and I expect the next iterations to give a clearer picture of how these mosquitoes respond to environmental and human-driven factors.

REFERENCES

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., … & Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504–507.

Google DeepMind, Google Research, & collaborators. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291. https://arxiv.org/abs/2507.22291

Kraemer, M. U. G., Sinka, M. E., Duda, K. A., Mylne, A. Q. N., Shearer, F. M., Barker, C. M., … & Hay, S. I. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife*, 4, e08347.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., … & Hay, S. I. (2014). Global spread of dengue virus types: mapping the 70-year history. *Trends in Microbiology*, 22(3), 138–146.

Messina, J. P., Brady, O. J., Pigott, D. M., Golding, N., Kraemer, M. U. G., Scott, T. W., … & Hay, S. I. (2015). The many projected futures of dengue. *Nature Reviews Microbiology*, 13(4), 230–239.

Tuanmu, M.-N., & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modeling. *Global Ecology and Biogeography*, 23(9), 1031–1045.

WorldPop. (2016). Asia Population Datasets, Version 2.0. Retrieved from https://www.worldpop.org/ DOI:10.5258/SOTON/WP00013