GeoHealth

RESEARCH ARTICLE

10.1029/2025GH001529

# A Machine Learning-Based Dynamic SST Index for Long-Lead Malaria Prediction in the Peruvian Amazon

**Mengxin Pan[1,2], Shineng Hu[1], Mark M. Janko[3], Benjamin F. Zaitchik[4], Ken Takahashi[5], Andres G. Lescano[6,7], Cesar V. Munayco[7,8], and William K. Pan[1,3]**

[1]Nicholas School of the Environment, Duke University, Durham, NC, USA, [2]Department of Geography, Simon Fraser University, Burnaby, BC, Canada, [3]Duke Global Health Institute, Duke University, Durham, NC, USA, [4]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA, [5]Instituto Geofísico del Perú, Lima, Peru, [6]Emerge, Emerging Diseases and Climate Change Research Unit, School of Public Health and Administration, Universidad Peruana Cayetano Heredia, Lima, Peru, [7]Clima, Latin American Center of Excellence for Climate Change and Health, Universidad Peruana Cayetano Heredia, Lima, Peru, [8]CDC Peru, National Center for Epidemiology, Disease Prevention and Control, Peruvian Ministry of Health, Lima, Peru

**Abstract** Malaria imposes a major health burden in the Peruvian Amazon, and its early warning is essential for effective disease prevention. The tropical sea surface temperature (SST) variability, fundamentally shaping the global weather patterns, may also alter malaria transmission and potentially improve its long-lead predictability. In this study, we propose a machine learning-based methodology that leverages comprehensive tropical SST variability for malaria prediction in the Peruvian Amazon. First, we demonstrate that significant correlations broadly exist between tropical SST anomalies and Peruvian malaria occurrence across different seasons and time lags, confirming the potential predictability from the tropical ocean. Then, we apply the self-organizing map to synthesize the spatiotemporally varying SST-malaria relationship and identify a unique dynamic SST index for Peruvian malaria. The dynamic SST index provides better performance (higher correlation coefficients and lower root mean square errors) in the generalized linear model, compared to the traditional El Niño–Southern Oscillation (ENSO) index, with lead times exceeding 3 months. Furthermore, the dynamic SST index captures the evolution of the ENSO life cycle from its precursor climate mode (Pacific Meridional Mode) and appears to influence Peruvian malaria by altering the local near-surface air temperature and specific humidity. Such underlying mechanisms provide the physically plausible basis for the long-lead predictability of Peruvian malaria using a machine learning-based remote predictor. Last but not least, we provide open-source code for broad applications in linking tropical SST variability and vector-borne disease transmission, or other climate-sensitive socioeconomic issues.

**Plain Language Summary** Malaria poses a serious health risk in the Peruvian Amazon, and its early warning system is vital for implementing effective prevention strategies. In this study, we explore the remote predictor for Peruvian malaria from the tropical ocean via a machine learning clustering algorithm (Self-organizing map; SOM). First, we demonstrate that significant correlations broadly exist between tropical sea surface temperature (SST) and Peruvian malaria occurrence across different seasons and time lags, indicating the potential predictive power from the ocean. Then, we identify a dynamic SST index by applying SOM to synthesize the complex SST-malaria relationship. Compared to the traditional El Niño–Southern Oscillation (ENSO) index, the dynamic SST index shows higher prediction performance in the single-predictor generalized linear model, with lead times exceeding 3 months. Moreover, we illustrate the underlying mechanism of how the dynamic SST index alters the local climate conditions and malaria transmission, providing the physically plausible basis for this data-driven remote predictor.

**Special Collection:**
Geospatial Artificial Intelligence (GeoAI) in Public Health

**Key Points:**

- Potential predictability for Peruvian malaria (significant correlations) broadly exists over the tropical ocean
- The dynamic sea surface temperature index, identified by self-organizing map, provides better prediction performance compared to conventional El Niño Southern Oscillation (ENSO) index
- It captures the development from Pacific Meridional Mode to ENSO and influences malaria by altering local air temperature and specific humidity

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

S. Hu and W. K. Pan,
shineng.hu@duke.edu;
william.pan@duke.edu

**Citation:**

Pan, M., Hu, S., Janko, M. M., Zaitchik, B. F., Takahashi, K., Lescano, A. G., et al. (2026). A machine learning-based dynamic SST index for long-lead malaria prediction in the Peruvian Amazon. *GeoHealth*, *10*, e2025GH001529. https://doi.org/10.1029/2025GH001529

Received 29 MAY 2025
Accepted 10 NOV 2025

**Author Contributions:**

**Conceptualization:** Mengxin Pan, Shineng Hu, William K. Pan
**Data curation:** Mengxin Pan
**Formal analysis:** Mengxin Pan
**Funding acquisition:** Shineng Hu, William K. Pan



## 1. Introduction

Vector-borne diseases, such as malaria, remain a substantial health burden worldwide, especially for vulnerable populations living in tropical and subtropical regions (World Health Organization, 2024). Effective prediction or early warning systems are crucial for malaria prevention by local public health authorities. Environmental factors, such as temperature and precipitation, directly influence the habitats of vectors (i.e., mosquitoes), thereby altering the malaria transmission dynamics (Liu et al., 2024; Paaijmans et al., 2009; Shapiro et al., 2017). This linkage makes malaria climate-sensitive and their prediction by environmental conditions has been achieved in many

**Investigation:** Mengxin Pan
**Methodology:** Mengxin Pan
**Project administration:** William K. Pan
**Resources:** William K. Pan
**Software:** Mengxin Pan
**Supervision:** Shineng Hu, William K. Pan
**Visualization:** Mengxin Pan
**Writing – original draft:** Mengxin Pan
**Writing – review & editing:** Mengxin Pan, Shineng Hu, Mark M. Janko, Benjamin F. Zaitchik, Ken Takahashi, Andres G. Lescano, Cesar V. Munayco, William K. Pan

studies (Haddawy et al., 2018; Janko et al., 2023; Wang et al., 2019). However, the prediction lead time by local environmental factors is usually limited to several weeks. On the other hand, the long-lead early warning system confers important advantages, such as enabling more comprehensive intervention planning, optimizing healthcare resource allocation, or increasing public engagement in epidemic prevention efforts. These practical benefits and needs inspire epidemiologists to search for remote predictors from the ocean, as the tropical ocean variability from several months ahead would modulate the continental weather pattern and alter the malaria transmission, thus potentially increasing the prediction lead time.

The primary remote predictor identified to date is the El Niño Southern Oscillation (ENSO) (Haines & Lam, 2023; Kovats et al., 2003). It denotes the irregular occurrence of anomalous warming and cooling sea surface temperature (SST) over the central and eastern tropical Pacific Ocean, known respectively as the El Niño and La Niña (McPhaden et al., 2020). ENSO events occur in the tropical Pacific, while they substantially influence the weather pattern worldwide through the processes known as climate teleconnections (Cai et al., 2020; Jong et al., 2021; Yeh et al., 2018). The primary index capturing the ENSO's phase and intensity is the averaged SST anomaly in the main ENSO region, such as the Niño 3.4 region (120–170°W, 5°S–5°N). This ENSO index has been widely and successfully used to predict various climate-sensitive vector-borne diseases in many regions, including malaria (Dhiman & Sarkar, 2017; Fisman et al., 2016; Lowe et al., 2017).

However, there are limitations to relying on the ENSO index as the only remote predictor, and the predictors for malaria could exist throughout the entire tropical ocean. For example, ENSO's predictive power for malaria in French Guiana and western India is limited or spatially inconsistent (Dhiman & Sarkar, 2017; Hanf et al., 2011). First, ENSO is the largest but not the only climatic fluctuation on the planet. The SST pattern in other ocean basins could also influence global weather patterns, either through interaction with ENSO or by directly influencing regional climates. Examples include the Pacific Meridional Mode (PMM), Indian Ocean dipole, and Atlantic Niño (Amaya, 2019; Cai et al., 2019; Deser et al., 2010; Li et al., 2016). Second, different types of ENSO events have been identified, and their impact varies significantly among different events or during different stages of the ENSO life cycle, making it difficult for a single ENSO index to consistently capture the full complexity of the relationship between ENSO and malaria transmission (Hu & Fedorov, 2018; Timmermann et al., 2018). Currently, people predominantly use well-established climate indices in their prediction models, rather than considering the variability over the entire tropical ocean (Bouma et al., 1997; Haines & Lam, 2023; Jalava et al., 2022; Kovats et al., 2003; Liyanage et al., 2022). In other words, the predictive potential from vast tropical oceans, which could exert substantial influence on the global climate conditions and malaria transmission dynamics, remains largely overlooked in global health applications.

In this study, we propose a framework to incorporate the SST variability over the entire tropical ocean in malaria prediction for the Peruvian Amazon. By applying a machine learning clustering algorithm (Self-organizing map; SOM), we synthesize the spatiotemporally varying SST-malaria relationship, then define a dynamic SST index as the unique predictor for Peruvian malaria. The dynamic SST index outperforms the conventional ENSO index in the single-predictor generalized linear model (GLM) with more than 3-month lead times. We further elucidate the underlying mechanisms linking pan-tropical ocean dynamics to climate conditions and malaria occurrence in the Amazon, providing a theoretical basis for the long-lead predictability. Moreover, we provide the open-source code and expect broad applications of our framework in linking tropical oceans and malaria in other regions or other climate-sensitive vector-borne diseases.

## 2. Methods

### 2.1. Data

The malaria data set comprises weekly counts of *Plasmodium vivax* cases in the Loreto region of the Peruvian Amazon from January 2000 to December 2022, obtained from the Center for Disease Control and Prevention (CDC) in Peru (CDC-Perú, 2023). The primary mosquito vector for *Plasmodium vivax* is *Anopheles darlingi* in the Peruvian Amazon. Loreto accounts for approximately 95% of national malaria occurrence and includes 49 districts, and the total number of *Plasmodium vivax* cases across 49 districts is used in our study. We use the malaria anomaly, defined as the observed number of malaria cases minus the long-term weekly mean during 2000–2022 (seasonality) for each epidemiological week, to isolate interannual variability. As shown by seasonality in Figure S1 in Supporting Information S1, the number of malaria cases in Loreto is perennial but exhibits

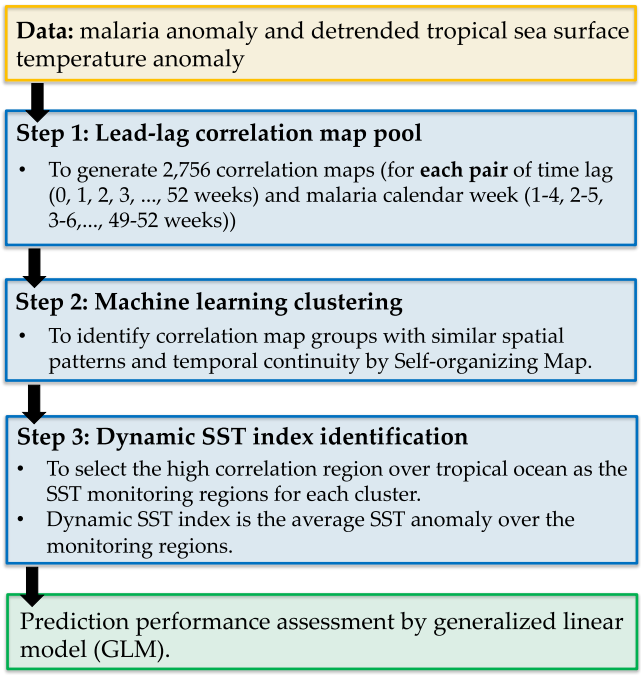


**Figure 1.** Flowchart of the "Correlation map—Self-organizing map—dynamic sea surface temperature index" framework for identifying potential predictors for climate-sensitive infectious disease (e.g., malaria in the Peruvian Amazon).

higher incidence during and immediately after the rainy season (approximately January through May, with peaks often in April–May).

For SST, we use the Optimum Interpolation SST version 2 (OISSTv2) from the National Oceanic and Atmospheric Administration, with 0.5° × 0.5° spatial and daily temporal resolution, from January 1999 to December 2022 (Huang et al., 2020). For each grid cell, we calculate the SST anomaly (SSTA) by subtracting the seasonally varying climatology from the raw SST value to remove the seasonal cycle. We aggregate daily anomalies into epidemiological weeks to align with the malaria data, then remove the linear trend in each calendar epidemiological week. The ENSO index is defined as the averaged detrended SSTA over the Niño 3.4 region (120–170°W, 5°S–5°N). To diagnose the underlying mechanism, we obtain the meteorological variables from the ERA5 reanalysis during 1999–2022 (Hersbach et al., 2019), including near-surface air temperature, near-surface specific humidity, total precipitation, mean tropospheric temperature, vertical motion at middle troposphere (500 mb), velocity potential and divergent wind at upper troposphere (200 mb), and vertical integral of atmospheric horizontal moisture transport. For observed local environmental conditions, we use the daily minimum and maximum temperatures from the CPC Global Unified Temperature data set (0.5° × 0.5°), and daily precipitation from CMORPH (0.25° × 0.25°) (Joyce & Xie, 2011).

### 2.2. "Correlation Map—Self-Organizing Map—Dynamic SST Index" Framework

In this study, we propose a three-step methodology to identify the dynamic SST index (the unique predictors from the tropical ocean) for long-lead prediction of malaria in the Peruvian Amazon. The flowchart of the three-step framework is provided in Figure 1. For broad applications of our methodology in linking tropical SST variability and vector-borne disease transmission, or other climate-sensitive socioeconomic issues, we also provide the open-source code on Zenodo (Pan, 2025).

The first step is constructing a pool of lead-lag correlation maps between the malaria anomaly and SSTA. We hypothesize that the SST-malaria relationship varies across seasons and time lags. Thus, we calculate a correlation map for each pair of SSTA calendar epidemiological week and malaria calendar epidemiological week, over all time lags from 0 to 53 weeks. In total, we generate 2,756 correlation maps, considering 52 calendar weeks and from 0 to 53 weeks' time lag. Only the SSTA in the tropical ocean (40°S–40°N) is considered, since the Peruvian Amazon is located in the tropics and the tropical ocean is the main driver of global teleconnection even for extratropical regions. To improve robustness, we aggregate the data into 4-week windows, which reduce noise and enhances the signal by increasing the sample size in the correlation calculation.

The second step is SOM clustering. We apply the SOM to cluster the 2,756 correlation maps into groups with similar spatial patterns. The main objective is to identify the unique SST index as the remote predictor for malaria in the Peruvian Amazon, which would be the averaged SSTA over regions with high malaria-SSTA correlation. However, the high correlation regions vary significantly across different time lags and calendar weeks (Figure 2), suggesting that the SST monitoring regions should not be static. On the other hand, selecting different monitoring regions for each week and time lag is impractical due to noisy data and limited sample sizes. Herein, we employ a machine learning clustering algorithm (SOM) to make a balance. SOM organizes samples in such a way that those in the same cluster are more similar to each other than to those in different clusters (Kohonen, 1998; Pan & Lu, 2020). Thus, we input the 2,756 correlation maps into the SOM to identify clusters with similar spatial correlation patterns. For dimension reduction, the SSTA fields are coarsened into 2° × 2° resolution. To emphasize the grids with high malaria-SST correlation, the grids with correlations smaller than 0.2 are set as 0. The 2 × 2 configuration (with 4 clusters) is selected. Sensitivity tests with alternate cluster numbers are provided in Supporting Information S1, showing that different choices of cluster number would not change the main result

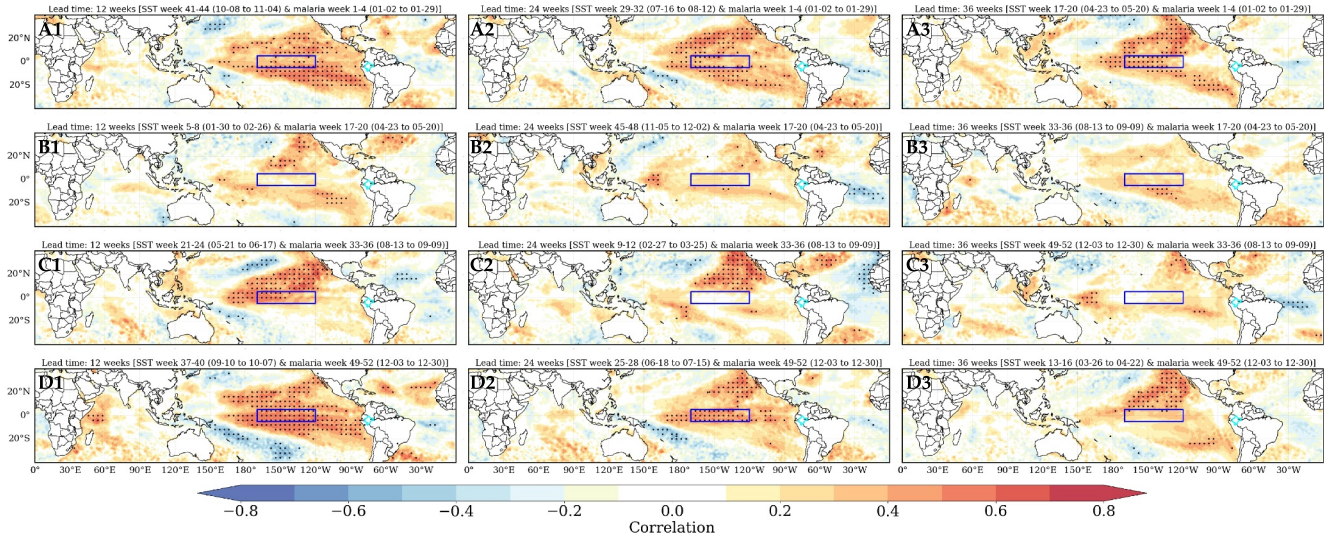


**Figure 2.** Correlation maps between malaria anomaly and SSTA fields in different calendar epidemiological weeks and time lags. Shading denotes the correlation coefficient ($r$-value), and scatters mark regions with significant correlations ($p$-value <0.001). Dark blue boxes represent the Niño 3.4 regions for the El Niño Southern Oscillation index calculation. Red stars mark our study region, the Peruvian Amazon (Loreto region in Peru). Estimated dates of each epidemiological week in 2000 are provided in the subtitles as a reference, while dates may vary slightly across years.

in this study. The choice of cluster number may vary when applying this methodology in different diseases over different regions.

The third step is dynamic SST index identification. After four SOM clusters are identified, an interesting finding is the temporal continuity of the clustering result, that is, SST-Malaria pairs in the same cluster share similar calendar weeks and time lags (Figure 3e). Thus, we select a constant SST monitoring region for pairs in a cluster to define the SST index and allow this region to vary across clusters. We define each SST monitoring region as the largest continuous region with high Malaria-SST correlation (i.e., the mean absolute value of correlation greater than 0.25 in the averaged correlation map in each cluster). As a result, a dynamic SST index is identified as the remote predictor for Peruvian malaria. The term "dynamic" implies that the SST monitoring regions change among different SOM clusters (i.e., shifts by season).

### 2.3. Generalized Linear Model for Predictor Comparison

To assess the predictability provided by the dynamic SST index in long-lead malaria prediction, we apply a generalized linear model (GLM) with a negative-binomial response. GLM is a log-linear regression model that extends ordinary linear regression to outcomes that are not Gaussian (e.g., count data). The outcome could follow a distribution from the exponential family. The infectious disease counts usually exhibit overdispersion (i.e., the variance larger than the mean), so we assume the number of malaria cases follows a negative binomial distribution (i.e., the Poisson variance with an extra dispersion parameter).

The $\text{Malaria}_t$ is the 4-weekly malaria occurrence in 2000–2022 and $\text{Malaria}_{\text{annual cycle}}$ is the mean malaria occurrence in each calendar 4-week from 2000 to 2022. We use the 4-week running mean to reduce the short-term noise of the malaria occurrence. The $Y_t$, which is the total malaria anomaly ($\text{Malaria}_t - \text{Malaria}_{\text{annual cycle}}$) after adding a constant, could be modeled as a negative binomial distribution with a mean parameter ($\mu_t$) and over-dispersion parameter ($\kappa_t$).

$$Y_t = \text{Malaria}_t - \text{Malaria}_{\text{annual cycle}} + \text{constant.} \quad (1)$$

$$Y_t \sim \text{Negative Binomial}\left(\mu_t, \kappa_t\right) \quad (2)$$

$$\log\left(\mu_t\right) = \alpha + \beta \times \text{SST\_index} \quad (3)$$

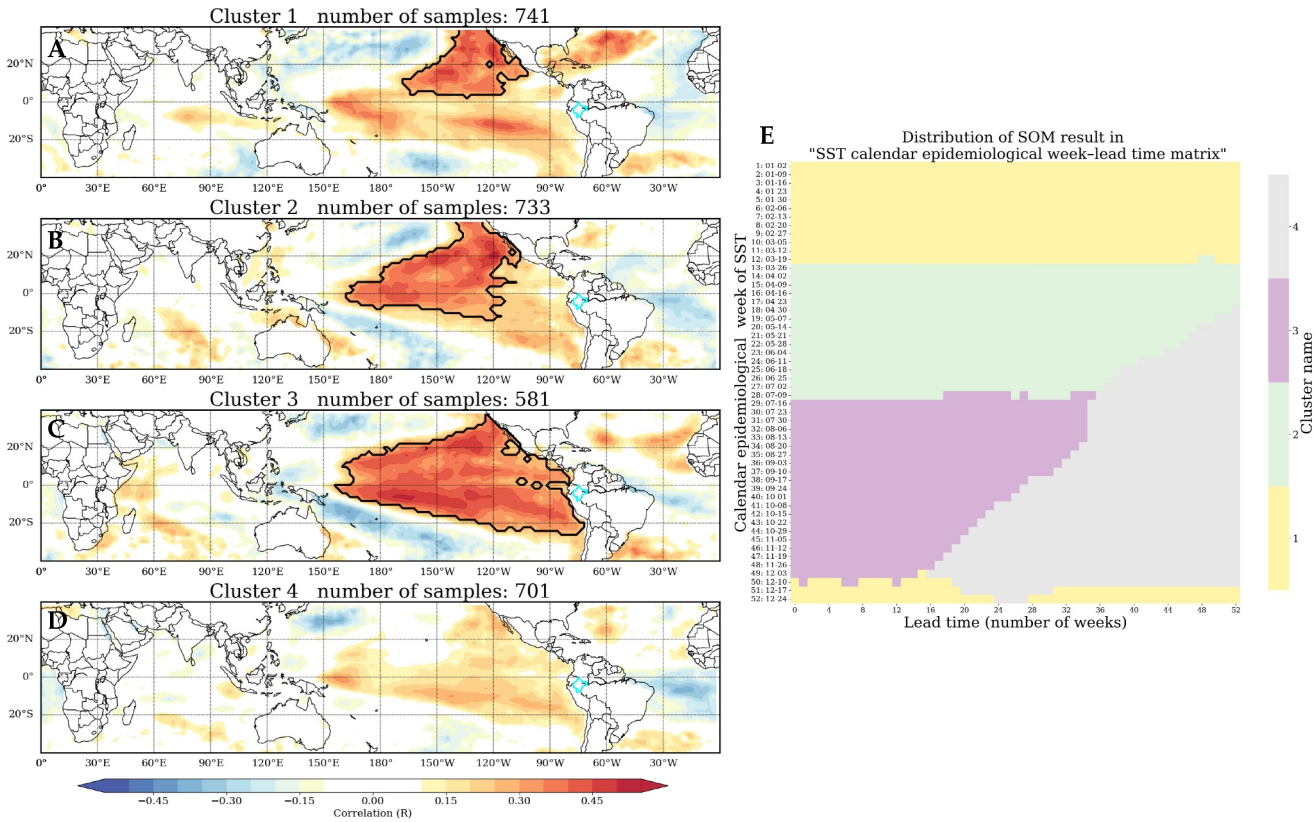


**Figure 3.** The self-organizing map (SOM) clustering results show the southwestward expansion of high-correlation regions. (a) to (d) the averaged correlation map of each SOM cluster, with black polygons highlighting the sea surface temperature monitoring regions with the high SST-malaria correlations. The number of samples in each cluster is provided in the subtitles. (e) The clustering results in the calendar epidemiological week—correlation lead time matrix, organized by SSTA's calendar epidemiological week in the *y*-axis.

We use the natural logarithm as the link function, assuming a linear relationship between the log of malaria anomalies and the dynamic SST index with intercept ($\alpha$) and slope ($\beta$). We fit separate models for each lead time, that is, SST index leads malaria from 4 to 54 weeks. For comparison, similar models are also constructed using the ENSO index. To prevent overfitting, we perform bootstrap cross-validation: in each of 10,000 iterations, we randomly select 80% of years for training (recomputing GLM coefficients) and test on the remaining 20%. We then compare models based on Pearson correlation and root mean square error (RMSE) of test-set predictions. A paired *t*-test evaluates whether the dynamic SST index yields significantly different performance from the ENSO index across iterations.

## 3. Results

### 3.1. High Correlation Between Peruvian Malaria and SSTA Outside the ENSO Region

We generated 2,756 correlation maps, considering 52 calendar weeks and from 0 to 53 weeks' time lag, to fully capture the spatiotemporally varying SST-malaria relationship. As illustrated in Figure 2, correlation maps reveal significant positive correlations between tropical SST anomalies (SSTA) and malaria anomalies in the Peruvian Amazon. Within the canonical ENSO region, SSTA and malaria co-vary strongly during certain seasons and time lags, consistent with known El Niño/La Niña impacts on Amazon climate and malaria transmission. However, in other cases, the strongest correlations emerge outside the ENSO region, such as over the subtropical northeast Pacific (Figure 2c2). The vanishing of significant correlation within the ENSO region suggests that relying solely on the static ENSO index may overlook predictability provided by other ocean basins. In other words, incorporating the entire tropical ocean is necessary and beneficial for better prediction of climate-sensitive vector-borne diseases, like malaria in the Peruvian Amazon.

### 3.2. Self-Organizing Map Synthesizes the SST-Malaria Relationship and Identifies the Dynamic SST Index

In addition, these correlation maps exhibit substantial spatial variability. For instance, when linking SSTA in weeks 17–20 (≈May) to malaria weeks 1–4 (≈January) at a 36-week lag, significant correlations appear across both tropical and subtropical Pacific (Figure 2a3). By contrast, correlating SSTA in weeks 5–8 (≈February) to malaria weeks 17–20 (≈May) at a 12-week lag yields nearly no significant correlations anywhere (Figure 2b1). This diversity and similarity of SST–malaria associations motivate our use of a machine-learning clustering algorithm (SOM) to synthesize the complex spatiotemporal patterns.

Applying SOM to the 2,756 correlation maps yields four distinct clusters (Figures 3a–3d) (Methods). The malaria-SST pairs in different clusters show distinct spatial patterns, each with unique high-correlation regions. When we plot the cluster membership in the calendar week versus lead-time matrix (Figure 3e), we see that maps within each cluster form continuous blocks. In other words, the malaria-SST pairs in the same cluster are temporally continuous, and clear boundaries can be found across different clusters. Notably, no date information was provided to the machine learning algorithm, so the temporal continuity of the clustering results underscores the robustness of the SOM approach here.

Moreover, the first three clusters all include the malaria-SST pairs with the same calendar SST weeks, but with different time lags (Figure 3e). Their high-correlation regions correspond to a seasonal progression: they extend southwestward from the southeast subtropical Pacific in early months (Cluster 1), toward the central tropical Pacific (Cluster 2), to almost the entire tropical Pacific later in the year (Cluster 3). This 3-stage expansion implies the development of ENSO from its precursor climate mode (PMM). The underlying mechanism of the PMM to ENSO progression and how they influence malaria transmission in the Peruvian Amazon will be illustrated in subsequent sections. We conduct sensitivity test for the SOM clustering in two ways. First, varying the number of SOM clusters consistently reproduces the PMM to ENSO progression, albeit broken into a different number of stages (Figures S2–S4 in Supporting Information S1). Second, a resampling sensitivity test—training the SOM on 75% of the years (18 out of 24) over four random iterations—yields similar three-stage structures in the calendar week versus lead-time matrix (Figure S5 in Supporting Information S1). These results confirm that the linkage from PMM-to-ENSO development to Peruvian malaria occurrence is a stable feature. The underlying mechanisms are detailed in the following sections.

The SOM result effectively summarizes the complex SST-malaria relationship and implies the necessity of the dynamic SST index for malaria prediction (i.e., shifting SST monitoring regions in different seasons), instead of the static ENSO index (i.e., monitoring the ENSO region consistently). Accordingly, we define the "dynamic SST index," with three seasonally varying monitoring regions (black polygons in Figures 3a–3c), to adapt to the seasonally shifting zone of highest malaria–SST correlation. Referring to Figure 3e, for SST calendar weeks falling into Cluster 1, we define the SST index as the averaged SSTA in the black polygon in Figure 3a. For SST calendar weeks falling into Cluster 2 and 3, we define the SST index by the black polygon in Figures 3b and 3c, respectively. Cluster 4 shows weak correlations globally, indicating a limited linkage between SST and Peruvian malaria occurrence with a large time lag. Thus, for the period in Cluster 4 (gray shading in Figure 3e), we adopt the high correlation region in Clusters 1 to 3 for the same SST weeks.

### 3.3. Dynamic SST Index Outperforms the ENSO Index in Malaria Prediction

We next compare the dynamic SST index with the traditional ENSO index in terms of correlation to malaria anomalies. As shown by the correlation matrix in Figure 4a, the dynamic SST index maintains significant positive correlations with malaria anomaly across nearly all seasons and time lags. Whereas, as shown by Figure 4b, the ENSO index has more insignificant and weaker correlations, particularly during Cluster 1 month. The difference matrix (Figure 4c) highlights that the dynamic SST index exhibits higher correlations with Peruvian malaria anomaly across all seasons, with the most pronounced improvement in January–March (Cluster 1 month). It suggests that during January to March, SSTA in the subtropical northeast Pacific plays a more essential role in altering Peruvian malaria than SSTA in the ENSO region.

To further quantify whether the dynamic SST index is a better predictor, we build generalized linear models (GLMs) for Peruvian malaria prediction, using the dynamic SST index and ENSO index as sole predictors, respectively (Methods). For both indices, GLMs are constructed for all prediction lead times (from 0 to 52 weeks).

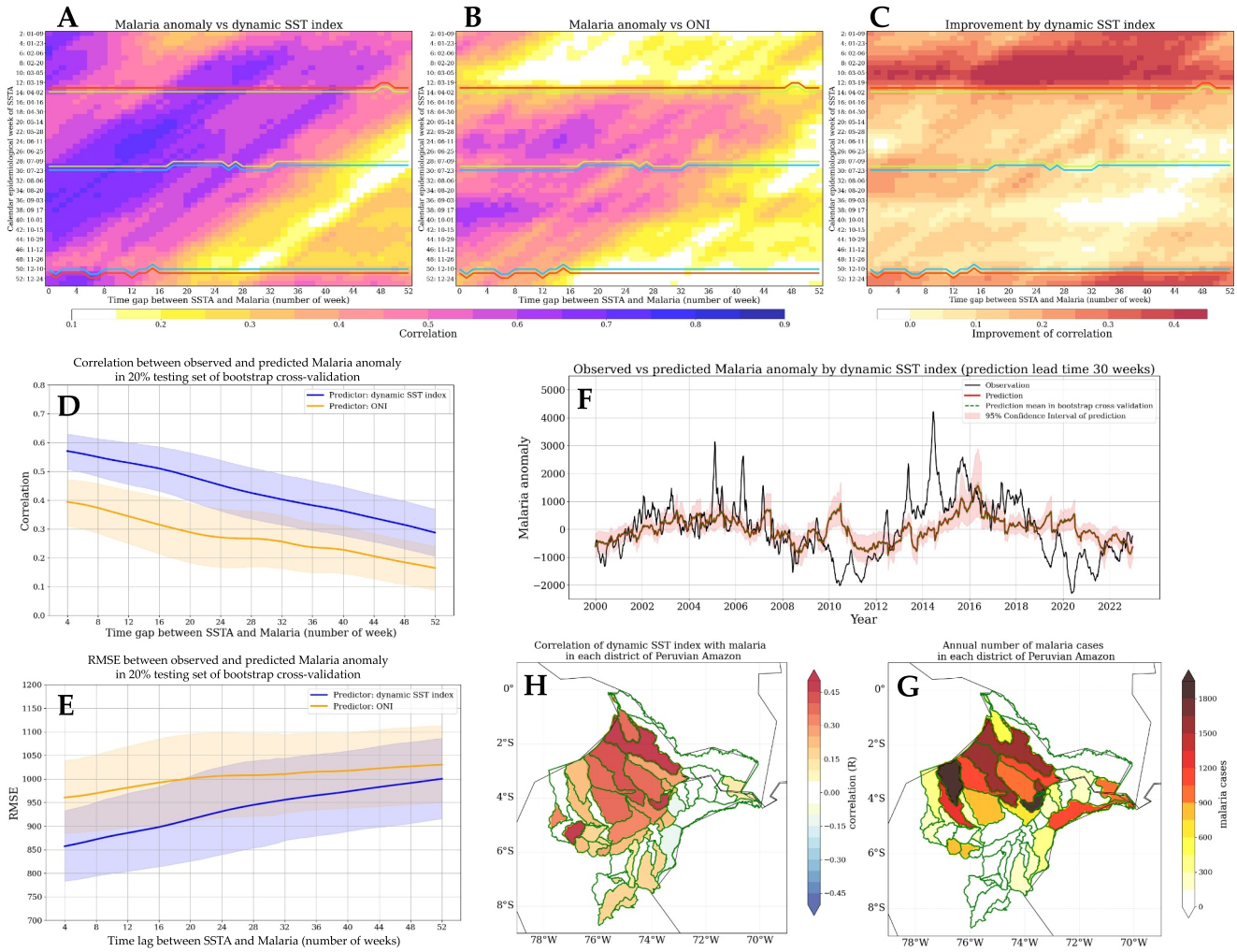


**Figure 4.** Comparison of the dynamic sea surface temperature (SST) index and El Niño Southern Oscillation (ENSO) index. (a) Correlation matrix between the dynamic SST index and malaria anomaly in the Peruvian Amazon in different SST calendar weeks and different lead times. We only show the significant correlation ($p$-value <0.001) here. Polygons with different colors mark 3 clusters, where Cluster 4 is replaced by Cluster 1–3 with the same SST week. (b) the same as (a) but for the ENSO index (c) Correlation matrix of dynamic SST index minus ENSO index ((a) minus (b)) (d) Correlation coefficient of the 20% testing set between observation and prediction in the generalized linear model (GLM) based on the dynamic SST index and ENSO, respectively. Shading shows the 90% confidence interval of the correlation coefficient in the 10,000 times bootstrap cross-validation (e) the same as (d) but for root mean square error. (f) Observed and predicted malaria anomaly with a 30-week lead time by the dynamic SST index. Red curves show the prediction by a model trained by all data, and red shadings show the confidence interval of GLM with an overdispersion parameter equal to 1. Green curves show the prediction mean of the testing set in the bootstrap cross-validation. (h) Correlation of dynamic SST index with malaria anomaly in each district over Peruvian Amazon with a 30-week lead time. (g) Annual occurrence of malaria in each district during 2000–2022.

We evaluate their out-of-sample skill via 20% test sets in the 10,000-iteration bootstrapping cross-validation. Figures 4d and 4e show that models with the dynamic index achieve higher median correlation and lower RMSE across all lead times. The paired $t$-test confirms these improvements are significant ($p < 0.001$). For example, when the dynamic SST index leads malaria anomaly for 30 weeks, the prediction by the dynamic SST index could relatively better capture the interannual variation of Peruvian malaria anomaly (Figure 4f). In the bootstrap cross-validation, the prediction means of testing sets (green line) are aligned with the prediction result based on the model trained by all data (red line). It demonstrates that the prediction results are not sensitive to the particular data set used for model fitting. We are also aware of the mismatch between the prediction and observation in many periods, such as the malaria outbreak period, the intense malaria intervention period (2006–2010), and the COVID-19 pandemic period (after 2019), implying the essential role of non-climatic factors in malaria transmission. We will discuss these caveats more in the discussion section. Moreover, in order to demonstrate the consistent positive correlation between dynamic SST index with malaria cases in different districts in Peruvian

10.1029/2025GH001529



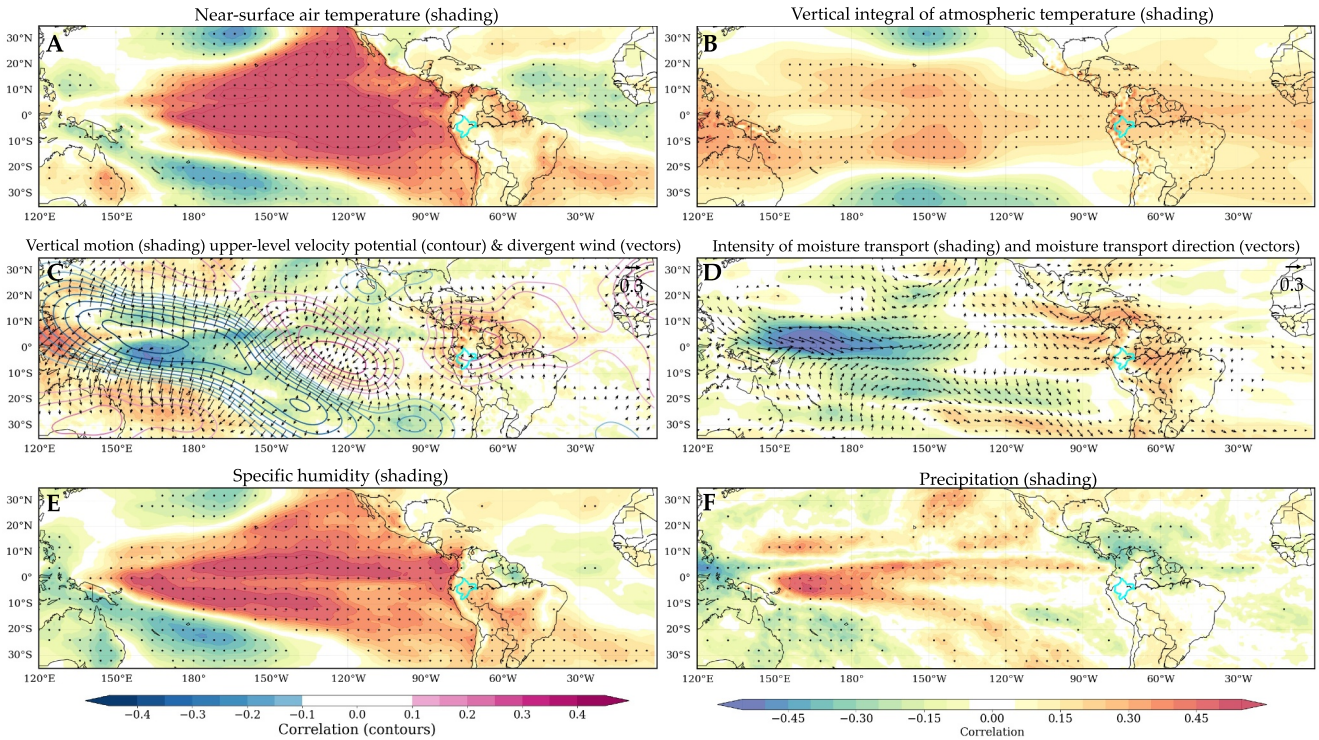


**Figure 5.** Lead-lag correlation maps between the dynamic sea surface temperature index and different tropical climate variables with 3-month lead times (a) Near-surface air temperature; (b) mean tropospheric temperature; (c) 500 mb vertical motion (shading) where positive value denotes descending motion, 200 mb velocity potential (contour) and divergent wind (vector) to represent the upper-level divergence/convergence of air mass, where the positive value of contour denotes convergence; (d) vertical integral of atmospheric horizontal moisture transport (shading), the vectors denote the relative correlation magnitude of zonal and meridional moisture transport intensity; (e) near-surface specific humidity; (f) total precipitation. The regions with significant correlation coefficient at 0.001 level are marked by black scatters.

Amazon, we provide the correlation maps across different districts (Figure 4h), along with the annual occurrence of malaria in each district (Figure 4g).

### 3.4. Underlying Mechanism of the Dynamic SST Index in Peruvian Malaria Prediction

In this section, we explore how the dynamic SST index influences malaria occurrence in the Peruvian Amazon from the climate dynamics perspective. As mentioned above, the three-stage southwestward expansion of the high correlation regions (Figures 3c–3e) well captures the development of ENSO from the PMM. The PMM, a subtropical-tropical interaction climate mode, is usually regarded as a precursor of ENSO from extratropical regions (Amaya, 2019). In boreal winter or spring, the SST warming of PMM initiates over coastal California (high correlation region of Cluster 1 in Figure 3a). In boreal summer, the SST warming extends southwestward to the central tropical Pacific due to the wind-evaporation-SST feedback (Cluster 2 in Figure 3b) (Shang-Ping Xie & Philander, 1994). The resulting SST warming and anomalous westerly wind over the central tropical Pacific could contribute to the development of El Niño events in the following autumn and winter (Cluster 3 in Figure 3c). The dynamic SST index well captures the entire life cycle of PMM to ENSO from the previous winter to the following winter, as shown by the correlation between the SST index and SSTA field in four seasons in Figure S6 in Supporting Information S1.

As mosquitoes and human beings live on the continents, they cannot feel the warming or cooling of the sea water over the Pacific. What they can feel is their surrounding environment. Thus, we further diagnose how the variation of Pacific SST, represented by the dynamic SST index, affects malaria transmission by altering the local climate conditions in Amazon. The dynamic SST index is positively correlated with near-surface air temperature with several months of lead time. The correlation maps with 3-month lead times are provided in Figure 5a, and the correlation maps from concurrent to 6-month lead times show similar patterns (Figure S6 in Supporting Information S1). The SST warming could warm up the entire tropical atmosphere by equatorial atmospheric Rossby

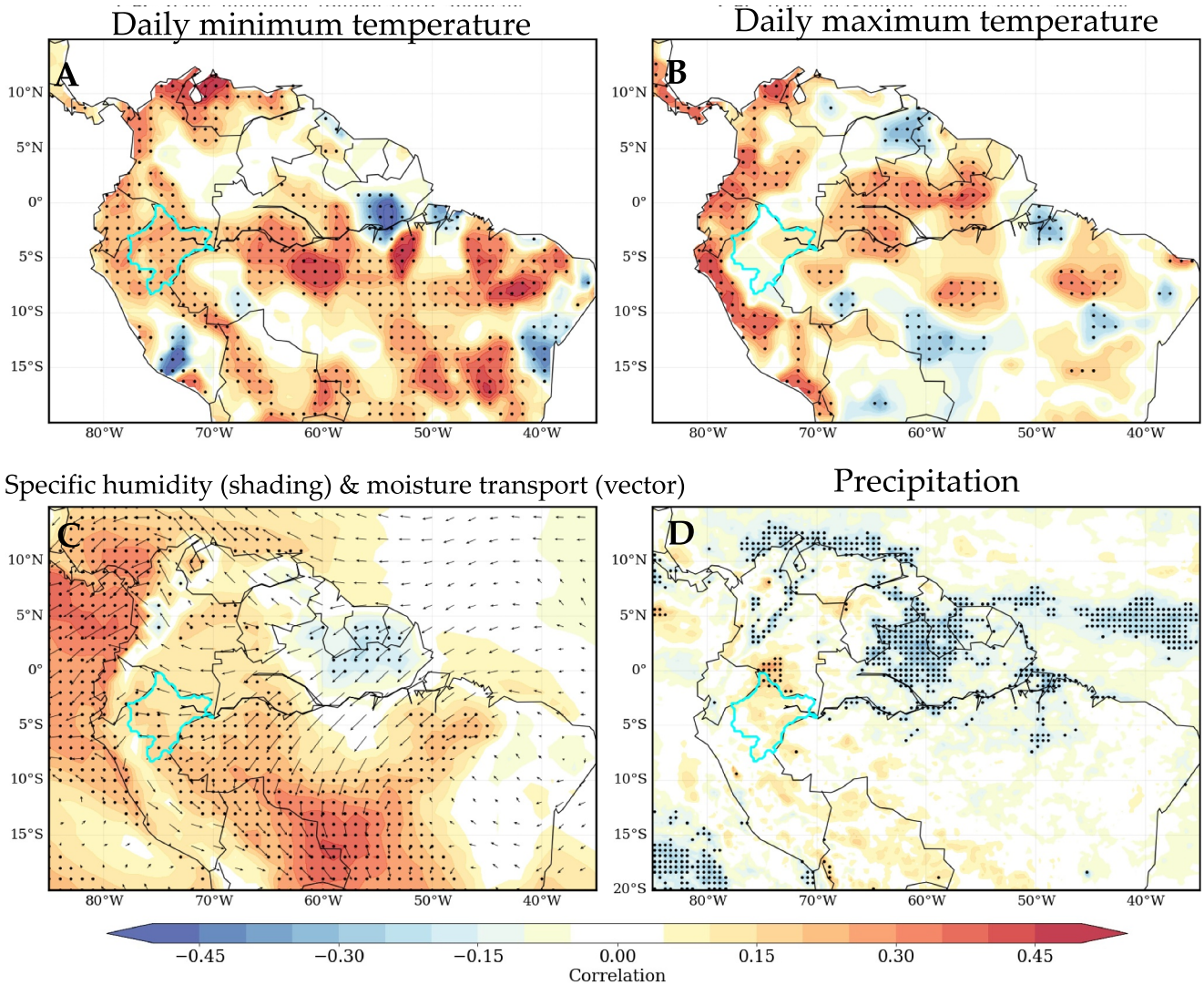


**Figure 6.** Concurrent correlation map between malaria anomaly and local environmental conditions. (a) Daily minimum temperature; (b) daily maximum temperature; (c) vertical integral of atmospheric horizontal moisture transport (shading), the vectors denote the relative correlation magnitude of zonal and meridional moisture transport intensity; (d) specific humidity (shading) and the same moisture transport vectors in (c); (e) precipitation from CMORPH. Light blue polygons mark the Peruvian Amazon.

and Kelvin waves (Figure 5b), which could increase the near-surface air temperature in the Amazon directly (Figure 5a). The warming effect lasts for the entire year. The increasing temperature would potentially facilitate the spread of malaria by providing favorable living conditions for mosquitos, supporting by the positive concurrent correlation between malaria anomaly and near surface air temperature, especially for the daily minimum temperature (Figures 6a and 6b).

At the same time, the tropical Pacific SST warming also enhances the moisture transport from the Atlantic into the Amazon (Figure 5d) and increases the near-surface specific humidity over the Amazon region (Figure 5e). The increasing moisture transport and humidity might also promote malaria transmission by providing more moist living conditions for mosquitos, supporting by the positive concurrent correlation between malaria anomaly and moisture transport and specific humidity (Figures 6c and 6d). The effect on precipitation is much weaker than other environmental conditions. On one hand, the enhanced moisture transport could enhance the precipitation. On the other hand, the SST warming over the central tropical Pacific will enhance the local convective rainfall and modulate the Pacific Walker Circulation, resulting in an anomalous descending motion in the Amazon region (Figure 5c) and suppressing the precipitation over Amazon (Figure 5f) (Cai et al., 2020). The two compensating effects lead to a weak net precipitation response over Peruvian Amazon. The correlation between malaria

anomaly and precipitation also shows weaker correlations than other climate variables over the entire Amazon region (Figure 6e). In summary, the warmer SST over the Pacific (positive dynamic SST index) leads to warmer and moister conditions over Amazon, which plausibly promote malaria transmission over the Peruvian Amazon.

Due to residential migration, environmental conditions in surrounding areas may also influence malaria occurrence in Peruvian Amazon. Therefore, we plot correlation maps between Peruvian malaria anomaly and environmental conditions across the broader northern South America. Indeed, significant correlations exist outside the Loreto. In summary, the underlying mechanism outlined above implies that the machine learning-based dynamic SST index captures the full PMM–ENSO cycle and its teleconnection to Amazon climate, providing a physically plausible basis for the long-lead malaria predictability demonstrated above.

## 4. Discussion

As the primary modulator of the global weather pattern, the tropical ocean contains various climate signals that might improve the prediction of malaria, or other climate-sensitive vector-borne disease worldwide. However, the vast ocean beyond the ENSO domain, which also exerts substantial influence on the global climate system and the disease transmission dynamics, is largely overlooked by current disease prediction models.

In this study, we go beyond the traditional approaches relying solely on the static climate indices and propose a machine learning-based framework that harnesses the variability of entire tropical ocean in the long-lead prediction of malaria in the Peruvian Amazon. First, we identified significant correlations between Peruvian malaria anomaly and SSTA across the vast tropical ocean, demonstrating predictive signals outside the commonly used ENSO domain. After synthesizing the spatiotemporally varying SST-malaria relationship by a machine learning clustering algorithm (SOM), we identify the dynamic SST index for Peruvian malaria. It provides lower RMSE and higher correlation than the conventional ENSO index in the single-predictor GLM model, in the paired *t*-test of the 10,000-iteration bootstrapping cross-validation. Furthermore, we demonstrate underlying mechanisms linking tropical ocean dynamics to local climate conditions and malaria occurrence, which provides the physically plausible basis for the long-lead malaria predictability. The dynamic SST index well captures the evolution of the ENSO life cycle from its precursor climate mode (PMM) from spring to winter. The Pacific SST warming, represented by the positive dynamic SST index, contributes to the surface air temperature warming and specific humidity increase in the Amazon, which are likely to promote malaria transmission by altering mosquitoes' habitat remotely.

While the dynamic SST index has shown improvements in predictive accuracy, several caveats still exist. First, predictions of single-predictor GLMs were less accurate during periods of intense disruptions, such as the 2006–2010 malaria control efforts and the COVID-19 pandemic (Janko et al., 2023). These discrepancies highlight the influence of non-environmental factors, such as public health interventions and socio-economic disruptions, which can buffer the impact of environmental predictors, especially for the remote predictors. Second, during outbreak periods, the model underestimated malaria occurrence, suggesting that human-driven factors may override environmental influences when disease transmission rates are unusually high. These non-environmental factors can also be confounders in a climate-driven prediction model: climate modes vary on characteristic timescales, and if a climate oscillation coincidentally aligns with a long-term variability or trend in relevant socio-economic or policy conditions, then predictive power might mistakenly be attributed to climate modes when in fact the driver was non-climatic. This caveat applies broadly to empirical modeling approaches, given the relative brevity of most infectious disease records relative to timescales of climate variability. We are aware that this study pays less attention to the potential confounders and focuses on searching for additional remote predictors for climate-sensitive disease from tropical SST. The role of human-driven factors and local environmental variabilities decides that the prediction of malaria by the ocean alone cannot be excellent. The model prediction skill could be further improved by incorporating additional socio-economic data, especially during these critical periods. Nevertheless, we find it promising that "PMM-ENSO" SST-based dynamic predictors show significant associations with both malaria occurrence and with physical mechanisms plausibly associated with malaria risk.

The broader implication of our study lies in its adaptability. Our framework bridges climate science and epidemiology and provides a new insight to improve the long-lead climate-sensitive vector-borne disease prediction by searching additional remote predictors from the tropical ocean. By scanning the full tropical SST field and deriving region- and disease-specific indices, the approach can improve both prediction accuracy and forecast horizon by several months. The application of our framework in Peruvian malaria suggests the transferability to

other vector-borne diseases and, potentially, other climate-sensitive socio-economic outcomes. We provide open-source code to support broad applications of this framework. As climate change continues to amplify climate-driven disruptions and exacerbate socioeconomic vulnerabilities, methodologies that connect climate dynamics to socioeconomic risk mitigation become increasingly critical for fostering global resilience in a warming world.

Moreover, we summarize some cautions that need to be taken during the applications. First, the study period needs to be long enough to capture the relationship between tropical SST variability and climate-sensitive outcomes. To reasonably capture the interannual climate-disease relation, a continuous record of longer than 20 years will probably be necessary. Second, we need to be aware of the potential influence of the non-environmental factors. Identifying the local socio-economic disruptions (e.g., vector control campaigns and migration) before applying this framework to find remote predictors will be helpful. Third, the interpretation of the SOM result is essential, including the underly mechanism of evolutionary spatial patterns and the temporal continuity of the calendar epidemiological week—correlation lead time matrix. The underlying rationale of long-lead predictability is important in forming a causal interpretation, which improves the reliability of the data-driven remote predictor.

## Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

## Data Availability Statement

Malaria surveillance data is publicly available from CDC-Peru (CDC-Perú, 2023). The OISSTv2 data is publicly accessible (Huang et al., 2020). The ERA5 reanalysis data are publicly accessible (Hersbach et al., 2019). The CPC global unified temperature is from https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html. The CMORPH precipitation is also publicly accessible (Joyce & Xie, 2011). An archived version of the code and data set of this study is available at Zenodo (Pan, 2025).

**Acknowledgments**
This project is supported by NIH (1R01AI151056, Pan PI), NASA (80NSSC22K1046, Pan PI) and the Dean's Research Venture Fund of Nicholas School of the Environment at Duke University (Pan, Hu co-PIs).

## References

Amaya, D. J. (2019). The Pacific meridional mode and ENSO: A review. In *Current Climate Change Reports, Current Climate Change Reports* (Vol. 5(4) pp. 296–307). Springer. https://doi.org/10.1007/s40641-019-00142-x

Bouma, M. J., Kovats, R. S., Goubet, S. A., Cox, J. S. H., & Haines, A. (1997). Global assessment of El Nino's disaster burden. *Lancet*, *350*(9089), 1435–1438. https://doi.org/10.1016/S0140-6736(97)04509-1

Cai, W., McPhaden, M. J., Grimm, A. M., Rodrigues, R. R., Taschetto, A. S., Garreaud, R. D., et al. (2020). Climate impacts of the El Niño–southern oscillation on South America. In *Nature Reviews Earth and Environment* (Vol. 1(4) pp. 215–231). Springer Nature. https://doi.org/10.1038/s43017-020-0040-3

Cai, W., Wu, L., Lengaigne, M., Li, T., McGregor, S., Kug, J. S., et al. (2019). Pantropical climate interactions. *Science*, *363*(6430), eaav4236. https://doi.org/10.1126/science.aav4236

CDC-Perú. (2023). Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (CDC-Perú). (CDC-Perú). *Ministerio de Salud, Peru. Malaria Case Surveillance Data (Loreto Region, Peru)*. Retrieved from https://www.dge.gob.pe/salasituacional/sala/index/salasit_dash/143

Deser, C., Alexander, M. A., Xie, S. P., & Phillips, A. S. (2010). Sea surface temperature variability: Patterns and mechanisms. In *Annual Review of Marine Science* Vol. 2(1), 115–143. https://doi.org/10.1146/annurev-marine-120408-151453

Dhiman, R. C., & Sarkar, S. (2017). El niño Southern oscillation as an early warning tool for malaria outbreaks in India. *Malaria Journal*, *16*(1), 122. https://doi.org/10.1186/s12936-017-1779-y

Fisman, D. N., Tuite, A. R., & Brown, K. A. (2016). Impact of El Niño southern oscillation on infectious disease hospitalization risk in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(51), 14589–14594. https://doi.org/10.1073/pnas.1604980113

Haddawy, P., Hasan, A. H. M. I., Kasantikul, R., Lawpoolsri, S., Sa-angchai, P., Kaewkungwal, J., & Singhasivanon, P. (2018). Spatiotemporal Bayesian networks for malaria prediction. *Artificial Intelligence in Medicine*, *84*(2018), 127–138. https://doi.org/10.1016/j.artmed.2017.12.002

Haines, A., & Lam, H. C. Y. (2023). El Niño and health in an era of unprecedented climate change. *The Lancet*, *402*(10415), 1811–1813. https://doi.org/10.1016/S0140-6736(23)01664-1

Hanf, M., Adenis, A., Nacher, M., & Carme, B. (2011). The role of El Niño southern oscillation (ENSO) on variations of monthly Plasmodium falciparum malaria cases at the cayenne general hospital, 1996-2009, French Guiana. *Malaria Journal*, *10*, 100. https://doi.org/10.1186/1475-2875-10-100

Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J. M., Nicolas, J., et al. (2019). Global reanalysis: Goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, *159*, 17–24. https://doi.org/10.21957/vf291hehd7

Hu, S., & Fedorov, A. V. (2018). Cross-equatorial winds control El Niño diversity and change. *Nature Climate Change*, *8*(9), 798–802. https://doi.org/10.1038/s41558-018-0248-0

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., et al. (2020). Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1. https://doi.org/10.1175/JCLI-D-20

Jalava, K., Dwivedi, S., Dhiman, R. C., Martineau, P., Ikeda, T., Minakawa, N., et al. (2022). Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning. *Frontiers in Public Health*, *10*, 962377. https://doi.org/10.3389/fpubh.2022.962377

Janko, M. M., Recalde-Coronel, G. C., Damasceno, C. P., Salmón-Mulanovich, G., Barbieri, A. F., Lescano, A. G., et al. (2023). The impact of sustained malaria control in the Loreto region of Peru: A retrospective, observational, spatially-varying interrupted time series analysis of the PAMAFRO program. *Lancet Regional Health - Americas*, *20*, 100477. https://doi.org/10.1016/j.lana.2023.100477

Jong, B. T., Ting, M., & Seager, R. (2021). Assessing ENSO summer teleconnections, impacts, and predictability in North America. *Journal of Climate*, *34*(9), 3629–3643. https://doi.org/10.1175/JCLI-D-20-0761.1

Joyce, R. J., & Xie, P. (2011). Kalman filter-based CMORPH. *Journal of Hydrometeorology*, *12*(6), 1547–1563. https://doi.org/10.1175/JHM-D-11-022.1

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1–3), 1–6. https://doi.org/10.1016/S0925-2312(98)00030-7

Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E., & Haines, A. (2003). El Niño and health. *Lancet (London, England)*, *362*(9394), 1481–1489. https://doi.org/10.1016/S0140-6736(03)14695-8

Li, X., Xie, S. P., Gille, S. T., & Yoo, C. (2016). Atlantic-induced pan-tropical climate change over the past three decades. *Nature Climate Change*, *6*(3), 275–279. https://doi.org/10.1038/nclimate2840

Liu, Q., Wang, Y., Deng, J., Yan, W., Qin, C., Du, M., et al. (2024). Association of temperature and precipitation with malaria incidence in 57 countries and territories from 2000 to 2019: A worldwide observational study. *Journal of Global Health*, *14*, 04021. https://doi.org/10.7189/JOGH.14.04021

Liyanage, P., Tozan, Y., Overgaard, H. J., Aravinda Tissera, H., & Rocklöv, J. (2022). Effect of El Niño–southern oscillation and local weather on Aedes dvector activity from 2010 to 2018 in Kalutara district, Sri Lanka: A two-stage hierarchical analysis. *The Lancet Planetary Health*, *6*(7), e577–e585. https://doi.org/10.1016/S2542-5196(22)00143-7

Lowe, R., Stewart-Ibarra, A. M., Petrova, D., García-Díez, M., Borbor-Cordova, M. J., Mejía, R., et al. (2017). Climate services for health: Predicting the evolution of the 2016 dengue season in Machala, Ecuador. *The Lancet Planetary Health*, *1*(4), e142–e151. https://doi.org/10.1016/S2542-5196(17)30064-5

McPhaden, M. J., Santoso, A., & Cai, W. (2020). El niño southern oscillation in a changing climate. In *Wiley-american geophysical union*.

Paaijmans, K. P., Read, A. F., & Thomas, M. B. (2009). Understanding the link between malaria risk and climate. Retrieved from www.pnas.orgcgidoi10.1073pnas.0903423106

Pan, M. (2025). The dataset and source code for "A Machine Learning-based Dynamic SST Index for Long-lead Malaria Prediction in the Peruvian Amazon" [Code and Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17117257

Pan, M., & Lu, M. (2020). East Asia atmospheric river catalog: Annual cycle, transition mechanism and precipitation. *Geophysical Research Letters*, *47*(15), e2020GL089477. https://doi.org/10.1029/2020GL089477

Shapiro, L. L. M., Whitehead, S. A., & Thomas, M. B. (2017). Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. *PLoS Biology*, *15*(10), e2003489. https://doi.org/10.1371/journal.pbio.2003489

Timmermann, A., An, S. I., Kug, J. S., Jin, F. F., Cai, W., Capotondi, A., et al. (2018). El Niño–southern oscillation complexity. *Nature*, *559*(7715), 535–545. https://doi.org/10.1038/s41586-018-0252-6

Wang, M., Wang, H., Wang, J., Liu, H., Lu, R., Duan, T., et al. (2019). A novel model for malaria prediction based on ensemble algorithms. *PLoS One*, *14*(12), 1–15. https://doi.org/10.1371/journal.pone.0226910

World Health Organization. (2024). *World malaria report 2024: Addressing inequity in the global malaria response*. World Health Organization. Retrieved from https://www.who.int/publications/i/item/9789240104440

Xie, S.-P., & Philander, S. G. H. (1994). A coupled ocean-atmosphere model of relevance to the ITCZ in the eastern Pacific. *Tellus, Series A*, *46 A*(4), 340–350. https://doi.org/10.3402/tellusa.v46i4.15484

Yeh, S. W., Cai, W., Min, S. K., McPhaden, M. J., Dommenget, D., Dewitte, B., et al. (2018). ENSO atmospheric teleconnections and their response to greenhouse gas forcing. *Reviews of Geophysics*, *56*(1), 185–206. https://doi.org/10.1002/2017RG000568