Supplementary information

Mapping the global distribution and environmental suitability for scrub typhus

Qian Wang[1,2], Tian Ma[3,4], Fangyu Ding[5,6], Ivo Elliott[2], Canjun Zheng[7], Nicholas Philip John Day[1,2], Benn Sartorius[2,8,9], Richard James Maude [1,2,10, *]

[1] Mahidol Oxford Tropical Medicine Research Unit (MORU), Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.
[2] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK.
[3] Yale Institute for Biospheric Studies, Yale University, New Haven, CT, United States.
[4] School of the Environment, Yale University, New Haven, CT, United States.
[5] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China.
[6] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China.
[7] Chinese Center for Disease Control and Prevention, Beijing, China
[8] Centre for Clinical Research (UQCCR), Faculty of Health, Medicine and Behavioural Sciences, University of Queensland, Brisbane, Australia.
[9] Department of Health Metric Sciences, School of Medicine, University of Washington, Seattle, USA.
[10] The Open University, Milton Keynes, UK
*Corresponding author

Contents

Table S1 GATHER (Guidelines for Accurate and Transparent Health Estimates Reporting) Checklist.

Item #	Checklist item	Reported on page #
Objectives and funding		
1	Define the indicator(s), populations (including age, sex, and geographic entities), and time period(s) for which estimates were made.	3-5
2	List the funding sources for the work.	14
Data Inputs		
For all data inputs from multiple sources that are synthesized as part of the study:		
3	Describe how the data were identified and how the data were accessed.	3
4	Specify the inclusion and exclusion criteria. Identify all ad-hoc exclusions.	SI 8
5	Provide information on all included data sources and their main characteristics. For each data source used, report reference information or contact name/institution, population represented, data collection method, year(s) of data collection, sex and age range, diagnostic criteria or measurement method, and sample size, as relevant.	SI 5-8
6	Identify and describe any categories of input data that have potentially important biases (e.g., based on characteristics listed in item 5).	14
For data inputs that contribute to the analysis but were not synthesized as part of the study:		
7	Describe and give sources for any other data inputs.	SI 8-11
For all data inputs:		
8	Provide all data inputs in a file format from which data can be efficiently extracted (e.g., a spreadsheet rather than a PDF), including all relevant meta-data listed in item 5. For any data inputs that cannot be shared because of ethical or legal reasons, such as third-party ownership, provide a contact name or the name of the institution that retains the right to the data.	SI 5-11
Data analysis		
9	Provide a conceptual overview of the data analysis method. A diagram may be helpful.	5-6
10	Provide a detailed description of all steps of the analysis, including mathematical formulae. This description should cover, as relevant, data cleaning, data pre-processing, data adjustments and weighting of data sources, and mathematical or statistical model(s).	5-6, SI 11-13
11	Describe how candidate models were evaluated and how the final model(s) were selected.	5-6, SI 13-14
12	Provide the results of an evaluation of model performance, if done, as well as the results of any relevant sensitivity analysis.	SI 13-14
13	Describe methods for calculating uncertainty of the estimates. State which sources of uncertainty were, and were not, accounted for in the uncertainty analysis.	6
14	State how analytic or statistical source code used to generate estimates can be accessed.	NA
Results and Discussion		
15	Provide published estimates in a file format from which data can be efficiently extracted.	NA
16	Report a quantitative measure of the uncertainty of the estimates (e.g. uncertainty intervals).	8
17	Interpret results in light of existing evidence. If updating a previous set of estimates, describe the reasons for changes in estimates.	13
18	Discuss limitations of the estimates. Include a discussion of any modelling assumptions or data limitations that affect interpretation of the estimates.	14

1. Scrub typhus data sources and data processing

1.1 National reported data

Occurrence data from national surveillance systems were acquired across five countries/regions: Mainland China, Japan, South Korea, Taiwan, and Thailand. The data were collected at varying spatial and temporal resolutions, case categories, and reporting criteria, as detailed below.

Mainland China: Scrub typhus is a vector-borne notifiable disease and has been included in the national surveillance system during 1952 to 1989 and from 2006 onwards in mainland China[1]. Physicians are required by law to report cases to the China Center for Disease Control and Prevention through the China Information System for Disease Control and Prevention (CISDCP). Scrub typhus case reports include basic demographic and clinical data including gender, age, occupation, residential address, date of onset of symptoms, laboratory diagnosis, and clinical outcome for each case. There was no information regarding the geographical distribution in the surveillance data from 1952 to 1989; the basic demographic and geographical information became available from 2006.

We obtained secured access to detailed data spanning from 2006 to 2019 through direct correspondence with relevant health authorities. To prevent revealing the identity of individuals, all the case data were anonymized prior to sharing. All scrub typhus cases were diagnosed and reported according to the diagnostic criteria issued by Chinese Center for Disease Control and Prevention, including suspect, clinically diagnosed, and laboratory-confirmed three categories. The diagnosis criteria and classification for those three categories are shown in Table S2.

Table S2. Scrub typhus diagnosis criteria and case classification in Mainland China.

Diagnosis criteria	
Epidemiological history	During the epidemic season, the patient has been in scrub typhus endemic areas within 3 weeks before the onset, and has a history of field activities, mainly including field work, rural fishing, camping training, sitting on grass, touching, and using straw, etc.
Clinical criteria	Fever; Lymphadenectasis; Rash; Specific eschar or ulceration
Confirmed criteria	-Positive Wail-Field test: single serum OXK titer ≥ 1:160; -Positive indirect immunofluorescence test IFA: paired serum IgG antibody titers increased by 4 times or more; -PCR nucleic acid test positive; -Pathogen isolated
Case classification	
Suspected case	-Epidemiology history + fever + lymphadenectasis or rash + clearly rule out other diseases -OR have fever + swollen lymph nodes + rash during the epidemic season
Clinical case	-Suspected case+ Specific eschar or ulceration -OR have epidemiological history + fever + characteristic eschar or ulceration
Confirmed case	-Suspected case + positive IFA or PCR or pathogen isolated; -OR clinical case + any confirmed criteria

In our study, we focused exclusively on clinical and confirmed cases, excluding all suspected cases from the national surveillance data. Included cases were geo-located based on the reported time and place of occurrence. Each clinical and confirmed case was meticulously mapped to its corresponding location, with spatial precision down to the county level.

Japan: The data used were obtained from the Infectious Diseases Statistics under the Infectious Disease Prevention Act (https://idsc.niid.go.jp/idwr/CDROM/Kako/NMenu.html) for 1999–2010, and the National Epidemiological Surveillance of Infectious Diseases (NESID) (https://www.niid.go.jp/niid/en/survaillance-data-table-english.html) for 2012 to 2022. In 1999, when infectious disease surveillance was placed under the NESID, scrub typhus was classified as a notifiable disease (compulsory reporting of all diagnosed cases)[2]. In addition to clinical manifestations that were indicative of rickettsial disease (e.g. fever, rash, and eschar), the respective case definitions for scrub typhus required a positive result from one or more of the following laboratory methods: rickettsial isolation; genome detection by polymerase chain reaction (PCR); serological evidence (IgM detection (≥ 1:80); or a ≥ 4-fold increase in the antibody titer between paired serum samples) by indirect fluorescent antibody (IFA) or indirect immunoperoxidase assay[3]. The weekly case number for individual prefectures from 1999 to 2023 were obtained.

South Korea: The Korea Centers for Disease Control and Prevention (KCDC) operate infectious disease surveillance systems to monitor national disease incidence. Since 1954, Korea has collected data on various infectious diseases in accordance with the Infectious Disease Control and Prevention Act[4]. All physicians (including those working in Oriental medicine) who diagnose a patient with an infectious disease or conduct a postmortem examination of an infectious disease case are obliged to report the disease to the system. These reported data are incorporated into the database of the National Infectious Disease Surveillance System, which has been providing web-based real-time surveillance data on infectious diseases since 2001 (https://www.kdca.go.kr/). The reporting data for scrub typhus in South Korea includes both suspected and confirmed cases. Suspected cases are identified as patients showing clinical symptoms (eschar, acute onset, rash, Lymphadenopathy and Hepatosplenomegaly) and having an epidemiological link suggestive of scrub typhus. Confirmed cases are those where the clinical symptoms are consistent with suspected cases and the infection should be verified through laboratory tests, such as isolation of the pathogen, a ≥ 4-fold rise in antibody titer, or detection of specific genes in clinical specimens. The data using in this study were collected at the district level, updated weekly, covering the period from 2001 to 2023.

Taiwan: An open infectious disease statistical data query system has been maintained by Taiwan CDC since 1996 to provide information on the number of confirmed cases of scrub typhus. Scrub typhus has been listed as a Category IV Notifiable Infectious Disease since 2007 in Taiwan based on the Communicable Disease Control Act. The publicly available online database, Taiwan National Infectious Disease Statistics System (TNIDSS), provided the number of confirmed cases of scrub typhus, the date of receipt, the date of onset, the date of diagnosis by the Department of Health, and the number of local or cases imported from overseas. Notification is defined as a suspected patient who meets clinical criteria (acute persistent high fever, headache, back pain, chills, night sweats, lymphadenopathy, painless eschar at the chigger bite and red skin with macules or papules after 1 week, sometimes accompanied by pneumonia or abnormal liver function). In addition, those who test positive in any one of the following tests are defined as positive[5]: (1) Clinical specimens (blood or skin wounds (eschar)) test positive for *O. tsutsugamushi* by nucleic acid detection; (2) Indirect immunofluorescene assay detects acute phase (or initial collection) serum, with a neutralization antibody titer of IgM of more than 1:80; and IgG titer was more than 1:320; (3) Using indirect immunofluorescence staining of matched (acute and convalescent) sera, a ≥4-fold increase in the IgG titer against *O. tsutsugamushi* is observed. The weekly confirmed case numbers at the district level in Taiwan from 2003 to 2023 were downloaded.

Thailand: We retrieved secured access to detailed data spanning from 2003 to 2022 through direct correspondence with relevant health authorities and obtained from the National Disease Surveillance System(R506). All scrub typhus cases were diagnosed and reported according to the diagnostic criteria issued by the Ministry of Public Health, including three categories of suspected, clinically diagnosed, and laboratory-confirmed. The diagnostic criteria and classification for those three categories are shown in Table S3. Cases were reported by governmental healthcare facilities including provincial hospitals, district hospitals and primary care units. Reporting from private healthcare facilities also occurred, albeit to a lesser extent[6]. Reported weekly case number from 2003 to 2022 were obtained at a district level and collated into a single dataset.

Table S3: Diagnostic criteria and reported classifications of scrub typhus in Thailand

Diagnosis criteria	
Clinical criteria	Acute febrile illness and an eschar with at least one other symptom including: - Headache - Myalgia - Arthralgia - Ocular or orbital pain - Petechial rash
Laboratory criteria	General findings suggestive of scrub typhus: - Low white cell count - Normal or low platelet count Disease-specific: - Detection of a four-fold rise in scrub typhus antibodies in paired sera by IFA *or* antibodies detected at a cut-off titre of ≥ 1:400 in a single sample *or*, - IIP obtained same result as IFA *or*, - *O. tsutsugamushi* PCR *or,* - culture positive from blood *or,* - Weil-Felix to OX-K with a titer of ≥ 1:320
Case classification	
Suspected case	Meets all clinical criteria and has a history of entering an area of grassland or forest.
Probable case	Fulfils clinical criteria and has general laboratory findings suggestive of scrub typhus *or* an epidemiological link to confirmed cases
Confirmed case	Fulfils clinical and any of disease-specific laboratory criteria

1.2 Data quality control

To ensure the reliability and consistency of the multi-source dataset used in our study, we implemented several rigorous data quality control measures. First, we cross-verified data from different sources, including published literature, public health surveillance data, and non-public surveillance data from certain countries. For data extracted from published literature, we cross-referenced it with national public health records to validate its accuracy. Additionally, we excluded data diagnosed solely by outdated or less specific diagnostic methods, such as the Weil-Felix test, or based only on clinical diagnoses without laboratory confirmation.

In terms of spatial and temporal data quality, we excluded records lacking specific spatial locations or those reported at broad administrative levels (e.g. administrative level 2 or above) or represented by large polygons exceeding 0.1x0.1 degrees to ensure that spatial information was detailed and precise. Similarly, records without explicit time information or with ambiguous temporal ranges were excluded. To maintain temporal consistency, records with extended operational times were transformed into yearly occurrences. Furthermore, we removed duplicate entries and conducted a thorough review of any outliers or inconsistent data points. Where necessary, data were either corrected or removed based on predefined criteria.

Lastly, we harmonized data across different sources by standardizing classifications and formats and recording sources. For non-public surveillance data, we applied additional scrutiny to verify authenticity and accuracy, cross-referencing these records with published literature and consulting with local experts when necessary. These steps were crucial in ensuring the overall integrity and accuracy of the dataset used in our analysis.

2. Explanatory covariates

We selected a suite of 28 covariates for inclusion in the model based on our systematic review findings and the availability of high-resolution spatial and temporal data. These covariates were chosen because they were shown to have a significant association with scrub typhus in our systematic review and are available as long-term data with a high spatial resolution of 5 km x 5 km and yearly or finer temporal resolution.

Table S4 Source of covariates.

Classification	Covariates	Resolution	Source
Climate	Minimum temperature	2.5 minutes, monthly	Worldclim
	maximum temperature		
	precipitation		
	Relative humidity	0.25 degree, monthly	ERA5
	Surface pressure		
	Wind speed		
Geographic	Elevation	2.5 minutes	Worldclim
	Normalized Difference Vegetation Index (NDVI)	0.05 degree, monthly	MODIS/Terra (MOD13A2)
	Enhanced Vegetation Index (EVI)		
	17 Landcovar	0.05 degree, annually	MODIS/Terra+Aqua (MCD12C1)
Socioeconomic	Population density	30 arc second, annually	WorldPop, Socioeconomic Data and Applications Center (SEDAC)
	Travel time to major cities (urban accessibility)	30 arc seconds	Weiss DJ, et.al[7]

Temperature and Precipitation: The WorldClim database (www.worldclim.org) consists of a freely available set of global climate data at a 2.5-minute spatial resolution which was compiled using weather data collected from world-wide weather stations. The monthly minimum temperature, maximum temperature, and precipitation data from 2000 to 2020 were obtained at a 2.5-minute spatial resolution, from which we generated annual information for each gridded cell at 0.05 degrees spatial resolution.

Elevation We obtained elevation data from the Shuttle Radar Topography Mission (SRTM) through the WorldClim database, originally at a 2.5-minute spatial resolution. Given that elevation is a static variable, we considered it to have no temporal resolution and assumed it remained constant over the nearly two-decade study period. To ensure consistency with other datasets used in our analysis, we resampled this elevation data to a 5kmx5km spatial resolution using the 'rasterio' package in Python.

Relative humidity: The ERA5 Essential Climate Variables dataset, available through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecv-for-climate-change?tab=overview), provides a selection of global climate data, including surface air relative humidity, at a 0.25° x 0.25° spatial resolution. This dataset is part of a comprehensive project aimed at assessing climate variability from 1979 to the present, with a focus on accuracy and temporal consistency across monthly to decadal time scales. The surface air relative humidity, expressed as a percentage, represents the ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water at the same temperature near the surface. The data, originally provided in GRIB format, were converted to raster TIFF format using the 'pygrib' package in Python. Subsequently, the data were resampled from the original 0.25-degree resolution to a 5 km x 5 km resolution using the 'gdal' package in Python, ensuring compatibility with our analysis framework.

Surface pressure and wind speed: The surface pressure and wind speed data used in this study were obtained from the ERA5 reanalysis project, accessible through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=form). ERA5 reanalysis combines model data with global observations into a consistent and complete dataset by utilizing data assimilation methods, where forecasts are periodically updated with new observations. The surface pressure data, available at a 0.25° x 0.25° spatial resolution, represent the atmospheric pressure at the Earth's surface, expressed in Pascals (Pa). This parameter measures the force per unit area exerted by the atmosphere at the surface, encompassing land, sea, and inland water. Surface pressure is often used in conjunction with temperature to calculate air density. The wind speed data is available at a 0.25° x 0.25° spatial resolution and is expressed in meters per second (m/s). This data provides information on the horizontal speed of air movement at a height of 10 meters above the Earth's surface and is a critical parameter in understanding surface wind patterns. While these data provide an averaged representation, actual wind observations can vary due to factors like local terrain, vegetation, and buildings. The eastward and northward components of the wind at 10 meters are also available. The original data in NetCDF format were processed to generate annual mean raster files using the 'netCDF4' package in Python, ensuring they are suitable for spatial analysis.

Landcover: The Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Climate Modelling Grid (CMG) (MCD12C1) Version 6.1 data product provides a spatially aggregated and reprojected version of the tiled MCD12Q1 Version 6 (https://doi.org/10.5067/MODIS/MCD12Q1.006) data product. We extracted global land cover data covering 17 different land use types from 2001 to 2020. The dataset is available at a spatial resolution of 0.05 degrees and a yearly temporal resolution and is provided in Hierarchical Data Format 4 (HDF4). The 17 land cover types follow the International Geosphere-Biosphere Programme (IGBP) classification scheme, and include: Water Bodies, Evergreen Needleleaf Forests, Evergreen Broadleaf Forests, Deciduous Needleleaf Forests, Deciduous Broadleaf Forests, Mixed Forests, Closed Shrublands, Open Shrublands, Woody

Savannas, Savannas, Grasslands, Permanent Wetlands, Croplands, Urban and Built-up Lands, Cropland/Natural Vegetation Mosaics, Permanent Snow and Ice, and Barren lands (Table S5). Each of these land cover types represents a distinct covariate in our analysis. We utilized Python and the 'GDAL' library to process and convert these data into 5 km raster formats, making them suitable for further spatial analysis in our study.

Table S5: MCD12C1 International Geosphere-Biosphere Programme (IGBP) legend and class descriptions.

Name	Value	Description
Water Bodies	0	At least 60% of area is covered by permanent water bodies.
Evergreen Needleleaf Forests	1	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
Evergreen Broadleaf Forests	2	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
Deciduous Needleleaf Forests	3	Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.
Deciduous Broadleaf Forests	4	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
Mixed Forests	5	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.
Closed Shrublands	6	Dominated by woody perennials (1-2m height) >60% cover.
Open Shrublands	7	Dominated by woody perennials (1-2m height) 10-60% cover.
Woody Savannas	8	Tree cover 30-60% (canopy >2m).
Savannas	9	Tree cover 10-30% (canopy >2m).
Grasslands	10	Dominated by herbaceous annuals (<2m).
Permanent Wetlands	11	Permanently inundated lands with 30-60% water cover and >10% vegetated cover.
Croplands	12	At least 60% of area is cultivated cropland.
Urban and Built-up Lands	13	At least 30% impervious surface area including building materials, asphalt, and vehicles.
Cropland/Natural Vegetation Mo- saics	14	Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.
Permanent Snow and Ice	15	At least 60% of area is covered by snow and ice for at least 10 months of the year.
Barren	16	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.
Unclassified	255	Has not received a map label because of missing inputs.

NDVI and EVI: We obtained the monthly vegetation indices grid data, including Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), from the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) for the period from February 2000 to December 2019. However, due to the absence of January 2000 data, only the data from 2001 to 2019 were processed and used in our analysis. These data are available as a gridded product at a spatial resolution of 0.05 degrees, in the sinusoidal projection. The NDVI serves as a continuity index with the existing National Oceanic and Atmospheric Administration-

Advanced Very High-Resolution Radiometer (NOAA-AVHRR) derived NDVI, providing consistency for long-term time series applications. The EVI, on the other hand, is designed to minimize canopy background variations and maintain sensitivity in regions with dense vegetation. EVI is particularly effective in high biomass areas and utilizes the blue band to correct for residual atmospheric contamination from smoke and sub-pixel thin clouds. Both NDVI and EVI are computed from surface reflectance that have been corrected for molecular scattering, ozone absorption, and aerosols. The original data were processed using Python and GDAL to generate raster formats suitable for further spatial analysis.

Population density We utilized population density data from the Gridded Population of the World, Version 4 (GPWv4). This dataset provides estimates of human population density (number of persons per square kilometre) for the years 2000, 2005, 2010, 2015, and 2020. The estimates are based on counts consistent with national censuses and population registers, adjusted to align with the 2015 Revision of the United Nations World Population Prospects (UN WPP) country totals. A proportional allocation gridding algorithm was used to distribute these population counts across 30 arc-second grid cells (~1 km at the equator), utilizing approximately 13.5 million national and sub-national administrative units. Additionally, we obtained population density data from WorldPop, adjusted to match individual country totals from the official United Nations population estimates. This data was derived from the corresponding Unconstrained individual countries 2000-2020 population count datasets, with population counts divided by pixel surface area to calculate density. The data was provided at a 30 arc-second resolution and was produced using the unconstrained top-down modelling method. For our analysis, we downloaded the 30 arc-second density raster data and resampled it to a 5kmx5km spatial resolution using the 'rasterio' package in Python, enabling integration with other spatial datasets.

Travel time to major cities (urban accessibility) We utilized a global map of urban accessibility, which provides land-based travel time to the nearest densely populated area for all regions between 85 degrees north and 60 degrees south, representing a nominal year of 2015 with a resolution of 30 arc seconds[7]. Densely populated areas are defined as contiguous regions with 1,500 or more inhabitants per square kilometre or areas with a majority of built-up land cover types coincident with a population center of at least 50,000 inhabitants. This map was produced through a collaboration between the University of Oxford Malaria Atlas Project (MAP), Google, the European Union Joint Research Centre (JRC), and the University of Twente, Netherlands, and was created using a combination of global datasets, including roads, railways, rivers, topography, and landcover types. Each pixel in the resulting accessibility map represents the modelled shortest travel time to a city. We resampled this map to a 5kmx5km spatial resolution using the 'rasterio' package in Python to ensure compatibility with other spatial datasets in our study.

3. Model

3.1 General overview

Model selection and parameter tuning are critical steps in developing predictive models, particularly when dealing with complex datasets and diverse modelling approaches. Proper model selection ensures that the chosen model accurately captures the underlying patterns in the data, while parameter tuning optimizes the model's performance by fine-tuning its internal settings. In our study, we employed a variety of methods to rigorously evaluate and select the best models for our analysis. The evaluation process included the use of cross-validation techniques, performance metrics such as Akaike Information Criterion AIC, Area Under the Curve (AUC), accuracy, and standard deviation, among others.

3.2 Generalized Additive Model (GAM)

The Generalized Additive Model (GAM) is a flexible statistical modeling technique that extends the traditional linear model by allowing non-linear relationships between the dependent and independent variables. Unlike linear models, which assume a straight-line relationship, GAMs enable each predictor to have its own smooth, non-linear function. This is particularly advantageous when dealing with complex data where relationships between variables are not strictly linear, allowing the model to capture more intricate patterns and dependencies.

GAMs are built by summing smooth functions of the predictors, which are estimated using techniques like splines or smoothing functions. The flexibility of GAMs comes from these smooth functions, which can adapt to the shape of the data, providing a more accurate fit.

We conducted a concurvity analysis to identify covariates with high concurvity—a condition akin to multicollinearity in linear models, which can obscure the true relationships between variables and lead to misleading results. Covariates with high concurvity were removed from the model to enhance clarity and interpretability (Figure S3). We identified several pairs of covariates exhibiting high concurvity, which is analogous to multicollinearity in linear models and can distort the interpretation of model coefficients. To address this, we systematically removed the covariates with high concurvity one by one from the model. After each removal, we compared the Akaike Information Criterion (AIC) to assess the impact on model performance. The final model excluded EVI, min temperature, savannas, urban-and-built-up and elevation and retained other covariates that contributed meaningfully to the predictive accuracy without introducing significant concurvity issues. To further refine the model and prevent overfitting, we applied a backward stepwise selection procedure. This process systematically removed less informative covariates, retaining only those that contributed the most to the model's predictive power.

Table S6: Concurvity Analysis and AIC Results for GAM.

Term1	Term2	Concurvity	AIC After Removal term1
s(Barren_or_Sparsely_Vegetated)	s(Savannas)	0.842696	38998.82
s(pop)	s(Urban_and_Builtup)	0.817263	38732.59
s(Savannas)	s(Barren_or_Sparsely_Vegetated)	0.842696	38720.48
s(evi)	s(ndvi)	0.956443	38828.73
s(ndvi)	s(evi)	0.956443	38850.57
s(Urban_and_Builtup)	s(pop)	0.817263	38730.09
s(tmax)	s(tmin)	0.951901	41302.29
s(tmin)	s(tmax)	0.951901	39713.45
s(elevation)	s(sp)	0.98785	39022.69
s(sp)	s(elevation)	0.98785	39093.96

3.3 Boosted Regression Trees (BRT) and Random Forest (RF) Model Selection

Boosted Regression Trees (BRT) is an advanced machine learning technique that combines the strengths of two powerful algorithms: boosting and decision trees. Boosting is an ensemble method that iteratively builds a series of decision trees, each one correcting the errors of its predecessor, to improve overall model accuracy. Decision trees, on their own, are simple and interpretable models, but when combined through boosting, they become a highly flexible and powerful predictive tool capable of capturing complex patterns in the data.

Random Forest (RF), another powerful machine learning algorithm, operates by constructing a multitude of decision trees during training. Unlike BRT, which builds trees sequentially, RF builds them independently and aggregates their outputs to make a final prediction. This ensemble approach helps reduce overfitting and improves the model's generalization to new data. RF is particularly valued for its robustness, accuracy, and ability to handle large datasets with numerous features.

In predictive modelling, particularly in the context of species distribution or disease occurrence, the presence-absence data balance is crucial. When the dataset is imbalanced, with significantly more absence cases than presence cases (or vice versa), it can bias the model, leading to poor predictive performance. In such scenarios, the model might simply predict the majority class (e.g. absence) most of the time, thereby failing to accurately identify true presence cases. To mitigate the potential bias from imbalanced data, we adopted a balanced dataset approach, ensuring an equal representation of presence and absence data. This was achieved by randomly selecting an equal number of absences from the entire absence dataset and combining them with the presence data to form the final dataset for modelling.

To optimize the performance of the BRT model, we considered two approaches for tuning parameters. The first approach involved using the caret package, which systematically explores all possible combinations of parameters based on the provided settings. This method requires extensive experience with the BRT model, as it offers a comprehensive exploration of the parameter space. However, for this analysis, we utilized the second approach, which involves the gbm.step function. This function is particularly effective in selecting the optimal number of trees for the model by focusing on the tree number that delivers the best performance, as indicated by the lowest cross-validation (CV) error. The gbm.step function allows for fine-tuning under pre-set values for other key parameters, such as tree complexity, learning rate, and bag fraction. The gbm.step function was employed to determine the most suitable number of trees. During this process, the other parameters of the algorithm were held at their default values: Tree Complexity: 4, Learning Rate: 0.005, Bag Fraction: 0.75 or 0.5 and Step Size: 500. Based on the result (Figure S4), we set 10000 as the trees' number. For RF, similar attention was given to balancing the presence-absence data and carefully tuning the hyperparameters. The default settings for RF were utilized with considerations for the number of trees and the maximum depth of each tree, optimizing the model's ability to generalize well to unseen data.

3.4 Relative importance of covariates

For the three sub-models (GAM, BRT, RF) and two stacking methods, the relative importance of covariates was assessed as follows: For GAM, the importance was determined by sequentially removing each covariate and comparing the change in model deviance. In contrast, both BRT and RF utilized their inherent importance functions to directly calculate the significance of each covariate. These analyses resulted in five distinct figures representing the relative importance of covariates across the models (Figure S5).

4.1 Result

4.1 Model performance

Table S7 summarizes the performance of the models assessed using both training and test datasets. The Random Forest (RF) model demonstrated the highest AUC in both the training (0.9999) and test datasets (0.9883), indicating its strong predictive performance. Additionally, two stacking ensemble models were evaluated: the RWM (Relative Weighted Mean) and CWM (Constrained Weighted Mean). The CWM model gave nearly perfect weighting to the RF sub-model, reflecting the RF model's dominance in prediction accuracy, with AUC values of 0.9856 and 0.9858 for training and test datasets, respectively. The RWM model showed more balanced weights across the sub-models, with AUC values slightly lower than those of the CWM model but still highly competitive. But based on the AUC value in the testing dataset, the RF model and its prediction were chosen to be our final model and result.

Table S7: Performance of models.

<table>
<tr><th>Model</th><th colspan="2">AUC training-5 fold</th><th>AUC test</th></tr>
<tr><td>GAM</td><td colspan="2">0.9438</td><td>0.9426</td></tr>
<tr><td>BRT</td><td colspan="2">0.9724</td><td>0.9718</td></tr>
<tr><td>RF</td><td colspan="2">0.9999</td><td>0.9883</td></tr>
<tr><td>Stack ensemble model-RWM</td><td colspan="2">0.9772</td><td>0.9775</td></tr>
<tr><td>Stack ensemble model-CWM</td><td colspan="2">0.9856</td><td>0.9858</td></tr>
<tr><th colspan="4">Weights of each sub-model in two stacking methods</th></tr>
<tr><td></td><td>GAM</td><td>BRT</td><td>RF</td></tr>
<tr><td>CWM</td><td>0.000001</td><td>0.000001</td><td>0.999998</td></tr>
<tr><td>RWM</td><td>0.2837838</td><td>0.3738739</td><td>0.3738739</td></tr>
</table>

4.2 Model robustness

As shown in Table S8, the robustness of the final RF model was extensively tested. The model maintained a high AUC across various scenarios, including different occurrence-to-absence ratios (e.g., 1:2, 1:5) and the use of different subsets of location data. The AUC remained above 0.98 in most scenarios, with consistently low prediction errors, as measured by the out-of-bag Brier score, consistently low. These results confirm the stability and reliability of the RF model under different conditions, further supporting its use for final predictions.

Table S8: Robustness test.

Model	**AUC test**	**Prediction. Error (out of bag Brier S.)**
RF (900 trees, null max.depth, 10 min node size, 5 mtry)	0.9883	0.040666625
RF (occurrence:absence=1:2)	0.9892	0.03395345
RF (occurrence:absence=1:5)	0.9913	0.02112522
RF use location (37929 records)	0.9553	0.07601925
RF (100 bootstraps)	Standard error: 0-0.02	

4.3 Area and population counting

Table S9: Predicted high environmental suitability areas and populations living in the environmental suitable areas of 215 countries.

Country	Area (km^2)	Population in millions (uncertainty)	Percentage of total population (uncertainty)	Local human scrub typhus occurrence
India	3231500	1166.32 (1160.26-1169.6)	98.7% (98.2%-98.9%)	Y
China	1330900	349.26 (301.13-427.75)	28.7% (24.7%-35.1%)	Y
Indonesia	1843375	223.89 (222.02-224.29)	97.9% (97.0%-98.0%)	Y
Brazil	8466600	175.88 (151.74-177.57)	96.8% (83.6%-97.8%)	N
Nigeria	924375	174.15 (173.68-174.15)	99.1% (98.8%-99.1%)	N
Pakistan	745975	166.44 (148.3-169.2)	93.8% (83.6%-95.4%)	Y
United States	1659100	150.06 (82.07-210.98)	53.5% (29.2%-75.2%)	N
Bangladesh	147275	139.46 (139.19-139.46)	99.1% (98.9%-99.1%)	Y
Mexico	1540975	89.77 (63.29-104.9)	79.5% (56.1%-92.9%)	N
Ethiopia	1091200	84.51 (64.29-90.98)	88.2% (67.1%-94.9%)	N
Egypt	527375	83.61 (53.82-86.54)	96.3% (62.0%-99.7%)	N
Vietnam	338050	80.33 (79.34-80.38)	99.2% (98.0%-99.2%)	Y
Zaire	2254100	76.58 (68.53-76.8)	99.6% (89.1%-99.9%)	N
Philippines	282150	75.93 (75.75-75.93)	94.1% (93.9%-94.1%)	Y
Thailand	531300	57.47 (57.29-57.47)	98.9% (98.6%-98.9%)	Y
Tanzania	931150	51.77 (46.08-52.25)	97.7% (87.0%-98.6%)	N
Sudan	2558600	51.2 (47.94-51.23)	99.9% (93.5%-99.9%)	N
Myanmar (Burma)	705525	46.14 (45.66-46.15)	97.5% (96.5%-97.6%)	Y
Kenya	579550	42.19 (34.29-42.68)	97.8% (79.5%-98.9%)	Y
South Korea	109625	40.87 (38.78-41.2)	96.4% (91.5%-97.2%)	Y
Uganda	242850	39.57 (36.19-39.87)	99.0% (90.6%-99.8%)	N
Colombia	1114000	33.92 (30.55-37.43)	81.3% (73.2%-89.7%)	N
Iraq	409975	32.9 (21.25-33.83)	93.3% (60.3%-95.9%)	N
South Africa	582850	30.94 (13.4-46.46)	62.9% (27.2%-94.5%)	N
Saudi Arabia	2014900	28.11 (21.4-28.75)	96.3% (73.3%-98.5%)	N
Venezuela	915975	27.96 (27.23-28.08)	98.0% (95.5%-98.5%)	N
Japan	137325	27.55 (10.14-62.7)	26.5% (9.8%-60.4%)	Y
Nepal	112000	26.77 (25.36-26.94)	98.3% (93.1%-98.9%)	Y
Malaysia	325100	26.46 (26.45-26.46)	96.8% (96.8%-96.8%)	Y
Mozambique	823525	26.42 (23.2-26.49)	97.1% (85.3%-97.4%)	N
Ghana	241025	25.32 (24.85-25.32)	98.4% (96.6%-98.4%)	N
France	261050	24.64 (8.11-41.38)	45.1% (14.8%-75.7%)	N
Yemen	431625	23.39 (15.75-24.68)	93.0% (62.6%-98.1%)	N
Angola	1263150	22.51 (18.61-22.62)	98.7% (81.6%-99.2%)	N
Madagascar	596125	21.8 (16.59-22.9)	90.8% (69.2%-95.4%)	N
Cameroon	469725	21.2 (20.64-21.22)	98.6% (96.0%-98.6%)	N
Niger	1246650	21.1 (20.65-21.1)	100.0% (97.9%-100.0%)	N
Germany	80850	20.9 (10.25-42.61)	30.4% (14.9%-62.0%)	N
Ivory Coast	317575	20.58 (20.44-20.58)	98.8% (98.1%-98.8%)	N
Burkina Faso	280800	18.54 (18.54-18.54)	100.0% (100.0%-100.0%)	N
Mali	1320750	17.67 (17.6-17.67)	100.0% (99.6%-100.0%)	N
Malawi	118225	17.24 (14.67-17.3)	99.6% (84.7%-99.9%)	N
Sri Lanka	65075	16.49 (16.42-16.49)	95.2% (94.8%-95.2%)	Y
Zambia	731625	15.62 (8.95-15.91)	98.1% (56.2%-100.0%)	N
Guatemala	111950	15.1 (13.91-15.41)	97.0% (89.4%-99.0%)	N
Argentina	697825	14.47 (4.68-28.25)	37.2% (12.0%-72.6%)	N
Zimbabwe	401525	14.36 (8.54-14.73)	96.8% (57.5%-99.3%)	N
Cambodia	186675	14.19 (14.09-14.19)	99.7% (99.0%-99.7%)	Y

Chad	1266500	14.07 (13.93-14.08)	99.9% (98.9%-100.0%)	N
Iran	567575	13.83 (9.83-25.49)	19.4% (13.8%-35.8%)	N
Senegal	202225	13.62 (10.66-13.62)	95.6% (74.8%-95.6%)	N
Morocco	165050	13.26 (3.04-22.09)	44.8% (10.3%-74.7%)	N
Syria	113100	12.9 (4.56-16.36)	72.7% (25.7%-92.2%)	N
Guinea	249350	11.9 (10.79-11.9)	98.7% (89.5%-98.7%)	N
Algeria	2074900	10.62 (2.68-20.1)	30.2% (7.6%-57.2%)	N
Ecuador	222100	10.55 (8.91-12.74)	72.0% (60.8%-86.9%)	N
Benin	118225	10.19 (9.57-10.19)	93.3% (87.6%-93.3%)	N
United Kingdom	96075	10.13 (0.77-34.41)	18.1% (1.4%-61.3%)	N
Burundi	25150	10.13 (6.7-10.84)	92.7% (61.3%-99.3%)	N
Rwanda	20900	10.1 (7.66-10.9)	88.9% (67.5%-95.9%)	N
Somalia	637950	10.07 (9.47-10.09)	97.5% (91.7%-97.8%)	N
North Korea	36700	10.05 (5.32-13.94)	46.7% (24.7%-64.7%)	N
Peru	885675	9.89 (7.19-16.12)	35.3% (25.6%-57.5%)	Y
Italy	38325	9.88 (3.26-23)	20.4% (6.7%-47.5%)	N
Turkey	76700	9.63 (3.31-24.53)	13.8% (4.7%-35.1%)	N
Haiti	27075	9.25 (9.16-9.26)	97.9% (96.9%-98.0%)	N
Cuba	111725	9.1 (9.07-9.1)	97.6% (97.3%-97.6%)	N
Dominican Republic	48825	8.43 (8.34-8.45)	87.4% (86.6%-87.6%)	N
Belgium	31800	8.12 (3.05-9.89)	79.5% (29.9%-96.8%)	N
United Arab Emirates	76650	7.98 (6.23-7.98)	99.0% (77.4%-99.0%)	N
Honduras	115750	7.36 (7.12-7.36)	99.0% (95.8%-99.0%)	Y
Papua New Guinea	452350	6.86 (6.3-6.91)	97.1% (89.2%-97.9%)	Y
Australia	6206675	6.62 (3.08-14.92)	30.9% (14.4%-69.6%)	Y
Laos	244075	6.34 (6.28-6.34)	100.0% (99.1%-100.0%)	Y
Spain	39200	6.13 (1.37-16.36)	15.9% (3.6%-42.5%)	N
Togo	58125	6.09 (6.08-6.09)	99.3% (99.1%-99.3%)	N
Paraguay	435725	5.89 (5.02-5.92)	99.6% (84.8%-100.0%)	N
Israel	16600	5.87 (4.26-6.24)	87.5% (63.5%-93.0%)	N
Nicaragua	130900	5.51 (5.4-5.51)	99.4% (97.4%-99.4%)	N
Bolivia	856100	5.45 (4.59-6.35)	54.1% (45.6%-63.1%)	N
Tunisia	115300	5.37 (2.01-8.31)	56.1% (21.0%-86.9%)	N
Sierra Leone	72725	5.23 (5.13-5.23)	98.9% (97.0%-98.9%)	N
El Salvador	20700	5.08 (5.08-5.08)	99.0% (99.0%-99.0%)	N
Netherlands	20250	5.06 (0.73-12.52)	34.5% (5.0%-85.4%)	N
Eritrea	123725	4.91 (3.67-5.1)	95.9% (71.8%-99.7%)	N
Central African Republic	627025	4.7 (4.43-4.7)	100.0% (94.2%-100.0%)	N
Chile	99025	4.68 (1.8-8.53)	30.5% (11.8%-55.6%)	Y
Congo	346225	4.62 (4.59-4.62)	99.1% (98.5%-99.1%)	N
Jordan	16025	4.49 (0.94-6.1)	61.0% (12.7%-82.8%)	N
Costa Rica	50825	4.14 (4.04-4.14)	98.8% (96.3%-98.9%)	N
Liberia	96400	4.06 (4.06-4.06)	95.9% (95.8%-95.9%)	N
Oman	327700	4.05 (3.58-4.09)	97.6% (86.2%-98.6%)	N
Mauritania	1112900	3.81 (2.29-3.81)	99.7% (60.0%-99.7%)	N
Singapore	425	3.68 (3.68-3.68)	78.7% (78.7%-78.7%)	N
Puerto Rico	8775	3.08 (3.08-3.08)	96.1% (95.8%-96.1%)	N
Libya	924800	3.08 (0.95-4.88)	53.9% (16.7%-85.5%)	N
Kuwait	18975	2.97 (2.39-2.97)	99.8% (80.6%-99.8%)	N
Panama	71525	2.9 (2.89-2.9)	92.4% (92.2%-92.4%)	N
Lebanon	4800	2.71 (1.97-3.12)	71.1% (51.6%-81.9%)	N
Afghanistan	44400	2.54 (1.12-6.33)	8.3% (3.6%-20.6%)	N

Canada	30600	2.53 (1.03-3.87)	8.1% (3.3%-12.3%)	N
New Zealand	144025	2.41 (0.71-3.41)	61.2% (18.1%-86.8%)	N
Austria	15675	2.39 (1.22-4.35)	33.7% (17.1%-61.2%)	N
Jamaica	11050	2.31 (2.3-2.31)	93.2% (92.9%-93.2%)	N
Botswana	624325	2.1 (1.85-2.11)	99.6% (87.5%-100.0%)	N
Namibia	607375	1.97 (1.34-2.19)	87.5% (59.5%-96.9%)	N
West Bank	4325	1.92 (0.67-2.61)	58.6% (20.4%-79.6%)	N
Qatar	11975	1.86 (0.4-1.86)	99.1% (21.4%-99.1%)	N
Croatia	28575	1.78 (0.64-2.61)	52.6% (19.0%-77.0%)	N
Hungary	27925	1.77 (0.55-5.11)	21.3% (6.7%-61.6%)	N
Serbia	16425	1.75 (0.79-3.42)	23.8% (10.8%-46.6%)	N
Gambia, The	11075	1.71 (1.36-1.71)	99.0% (78.7%-99.0%)	N
Poland	6925	1.6 (0.44-10.5)	4.9% (1.3%-31.8%)	N
Gabon	261900	1.53 (1.52-1.53)	99.7% (99.3%-99.7%)	N
Guinea-Bissau	32325	1.5 (1.12-1.5)	88.1% (65.8%-88.1%)	N
Gaza Strip	500	1.35 (1.13-1.35)	100.0% (83.6%-100.0%)	N
Slovakia	12000	1.33 (0.69-3.01)	28.0% (14.4%-63.5%)	N
Ireland	54875	1.18 (0.04-3.04)	29.2% (0.9%-75.5%)	N
Swaziland	18750	1.16 (0.91-1.18)	97.8% (76.8%-99.3%)	N
Trinidad and Tobago	4600	1.02 (1.02-1.02)	97.4% (97.4%-97.4%)	N
Mauritius	1900	0.96 (0.95-0.96)	95.1% (94.2%-95.1%)	N
Slovenia	7725	0.85 (0.41-1.39)	48.3% (23.6%-79.3%)	N
Cyprus	7950	0.81 (0.47-0.91)	83.9% (48.6%-93.6%)	N
Equatorial Guinea	26550	0.79 (0.78-0.79)	97.7% (97.1%-97.7%)	N
Djibouti	21850	0.78 (0.67-0.78)	99.9% (85.3%-99.9%)	Y
Portugal	6425	0.68 (0.2-3.73)	8.4% (2.4%-45.8%)	N
Bhutan	20275	0.61 (0.59-0.73)	68.3% (66.1%-82.0%)	Y
Romania	17425	0.6 (0.19-5.44)	3.8% (1.2%-34.5%)	N
Ukraine	6475	0.59 (0.26-5.31)	1.6% (0.7%-14.2%)	N
Uruguay	8975	0.55 (0.16-2.21)	17.7% (5.3%-71.5%)	N
Western Sahara	257825	0.55 (0.28-0.55)	93.0% (47.2%-93.0%)	N
Guyana	212500	0.55 (0.54-0.55)	96.5% (95.9%-96.5%)	N
Comoros	1425	0.54 (0.51-0.54)	82.6% (78.7%-82.6%)	N
Czech Republic	4500	0.54 (0.13-3.47)	5.9% (1.5%-38.2%)	N
Fiji	16225	0.53 (0.53-0.53)	79.5% (79.3%-79.5%)	N
Norway	12500	0.52 (0.18-1.23)	12.2% (4.2%-28.9%)	N
Reunion	1875	0.52 (0.44-0.56)	77.3% (65.9%-82.6%)	N
Greece	6775	0.49 (0.14-2.38)	5.5% (1.5%-26.9%)	N
Suriname	146125	0.48 (0.48-0.48)	99.8% (99.7%-99.8%)	N
Bahrain	525	0.46 (0.23-0.46)	83.8% (41.6%-83.8%)	N
Brunei	5850	0.37 (0.37-0.37)	97.1% (97.1%-97.1%)	N
Luxembourg	1575	0.37 (0.13-0.51)	70.1% (25.2%-96.4%)	N
Russia	6125	0.35 (0.13-4.59)	0.3% (0.1%-3.8%)	Y
Belize	22325	0.33 (0.33-0.33)	97.5% (96.7%-97.5%)	N
Cape Verde	3125	0.29 (0.24-0.31)	85.3% (71.1%-90.6%)	N
Guadeloupe	1525	0.28 (0.28-0.28)	90.3% (90.3%-90.3%)	N
Malta	225	0.27 (0.24-0.27)	82.1% (73.8%-82.1%)	N
Macau	25	0.25 (0.25-0.25)	100.0% (100.0%-100.0%)	N
French Guiana	84100	0.24 (0.24-0.24)	99.0% (99.0%-99.0%)	N
Martinique	975	0.24 (0.24-0.24)	86.8% (86.8%-86.8%)	N
Solomon Islands	22075	0.24 (0.24-0.24)	77.7% (77.6%-77.7%)	Y
Moldova	700	0.23 (0.14-1.21)	6.8% (4.1%-35.4%)	N
Azerbaijan	2050	0.19 (0.06-1.51)	2.1% (0.6%-17.1%)	N

Bahamas, The	9375	0.18 (0.17-0.18)	76.2% (72.1%-76.8%)	N
French Polynesia	1500	0.16 (0.14-0.16)	80.1% (72.8%-80.1%)	N
Western Samoa	2400	0.15 (0.15-0.15)	93.5% (92.9%-93.5%)	N
Bosnia and Herzegovina	2425	0.15 (0.03-0.9)	4.7% (0.8%-28.0%)	N
New Caledonia	18225	0.15 (0.14-0.15)	79.9% (73.3%-80.2%)	N
Barbados	375	0.14 (0.14-0.14)	94.2% (94.2%-94.2%)	N
Sao Tome and Principe	850	0.13 (0.13-0.13)	92.6% (92.6%-92.6%)	Y
Guam	475	0.12 (0.12-0.12)	80.2% (80.2%-80.2%)	N
Georgia	700	0.12 (0.02-0.32)	3.4% (0.7%-9.5%)	N
Mayotte	275	0.12 (0.12-0.12)	80.1% (80.1%-80.1%)	N
Vanuatu	10000	0.11 (0.11-0.11)	78.1% (77.6%-78.1%)	Y
Bulgaria	975	0.1 (0.02-1.4)	1.7% (0.3%-23.8%)	N
Antigua and Barbuda	375	0.06 (0.06-0.06)	97.3% (96.9%-97.3%)	N
Dominica	750	0.06 (0.06-0.06)	95.4% (95.4%-95.4%)	N
Switzerland	175	0.05 (0.01-2.38)	0.7% (0.2%-32.0%)	N
Grenada	275	0.05 (0.05-0.05)	87.3% (87.3%-87.3%)	N
St. Vincent and the Grenadines	300	0.05 (0.05-0.05)	98.2% (95.0%-98.2%)	N
Seychelles	125	0.05 (0.05-0.05)	69.7% (69.7%-69.7%)	N
Sweden	175	0.04 (0-1.78)	0.5% (0.0%-21.8%)	N
Tonga	275	0.04 (0.04-0.04)	69.9% (68.9%-69.9%)	N
Man, Isle of	225	0.03 (0-0.05)	33.4% (0.0%-59.0%)	N
Federated States of Micronesia	350	0.02 (0.02-0.02)	71.6% (71.6%-71.6%)	N
Jersey	50	0.02 (0.01-0.04)	26.5% (6.6%-49.1%)	N
St. Lucia	75	0.02 (0.02-0.02)	13.4% (13.4%-13.4%)	N
Tajikistan	150	0.02 (0-1.3)	0.2% (0.0%-16.0%)	N
St. Kitts and Nevis	150	0.02 (0.01-0.02)	62.8% (52.3%-62.8%)	N
American Samoa	50	0.02 (0.02-0.02)	45.9% (45.9%-45.9%)	N
Virgin Islands	175	0.02 (0.02-0.02)	73.6% (73.6%-73.6%)	N
Northern Mariana Islands	125	0.02 (0.02-0.02)	81.3% (81.3%-81.3%)	N
Albania	125	0.01 (0-0.42)	0.3% (0.0%-18.7%)	N
British Virgin Islands	25	0.01 (0.01-0.01)	40.5% (40.5%-40.5%)	N
Uzbekistan	50	0 (0-0.89)	0.0% (0.0%-3.3%)	N
Pacific Islands (Palau)	250	0 (0-0)	71.0% (71.0%-71.0%)	Y
Denmark	50	0 (0-1.14)	0.1% (0.0%-25.7%)	N
Montserrat	100	0 (0-0)	100.0% (100.0%-100.0%)	N
Cayman Islands	100	0 (0-0)	3.9% (3.9%-3.9%)	N
Faroe Islands	50	0 (0-0.02)	2.3% (0.0%-56.7%)	N
Turks and Caicos Islands	100	0 (0-0)	43.6% (43.6%-43.6%)	N
Armenia	0	0 (0-0.01)	0.0% (0.0%-0.0%)	N
Andorra	0	0 (0-0)	0.0% (0.0%-15.2%)	N
Byelarus	0	0 (0-0.58)	0.0% (0.0%-7.0%)	N
Estonia	0	0 (0-0.36)	0.0% (0.0%-32.0%)	N

Finland	0	0 (0-0.31)	0.0% (0.0%-6.8%)	N
Guernsey	0	0 (0-0)	0.0% (0.0%-66.3%)	N
Greenland	0	0 (0-0.02)	0.0% (0.0%-0.0%)	N
Iceland	0	0 (0-0)	0.0% (0.0%-1.5%)	N
Jan Mayen	0	0 (0-0)	0.0% (0.0%-6.7%)	N
Kyrgyzstan	0	0 (0-0.07)	0.0% (0.0%-0.2%)	N
Kazakhstan	0	0 (0-0.01)	0.0% (0.0%-0.5%)	N
Latvia	0	0 (0-0.35)	0.0% (0.0%-22.4%)	N
Lithuania	0	0 (0-0.99)	0.0% (0.0%-19.3%)	N
Liechtenstein	0	0 (0-0)	0.0% (0.0%-0.0%)	N
Lesotho	0	0 (0-0.46)	0.0% (0.0%-50.5%)	N
Mongolia	0	0 (0-0.05)	0.0% (0.0%-0.0%)	N
Macedonia	0	0 (0-0)	0.0% (0.0%-2.8%)	N
Monaco	0	0 (0-0)	0.0% (0.0%-0.0%)	N
Montenegro	0	0 (0-0.05)	0.0% (0.0%-8.7%)	N
St. Pierre and Miquelon	0	0 (0-0.01)	0.0% (0.0%-46.6%)	N
San Marino	0	0 (0-0)	0.0% (0.0%-38.5%)	N
Svalbard	0	0 (0-0)	0.0% (0.0%-0.1%)	N
Turkmenistan	0	0 (0-0.38)	0.0% (0.0%-7.4%)	N

Table S10: Countries/regions where population weighted environmental suitability > 0.5 but no confirmed cases in humans have been reported.

GMI code	Country	Local human scrub typhus occurrence	Environmental suitability (population weighted)
ATG	Antigua and Barbuda	N	0.8431415
AGO	Angola	N	0.7060163
BHR	Bahrain	N	0.5935044
BRB	Barbados	N	0.8235087
BWA	Botswana	N	0.6097312
BEL	Belgium	N	0.5446713
BHS	Bahamas, The	N	0.6562729
BLZ	Belize	N	0.8506504
BOL	Bolivia	N	0.5261988
BEN	Benin	N	0.8186688
BRA	Brazil	N	0.7707489
BRN	Brunei	N	0.9156949
BDI	Burundi	N	0.7044257
TCD	Chad	N	0.8026985
COG	Congo	N	0.8631865
ZAR	Zaire	N	0.8219726
CMR	Cameroon	N	0.8615318
COM	Comoros	N	0.7199026
COL	Colombia	N	0.7266011
MNP	Northern Mariana Islands	N	0.7243235
CRI	Costa Rica	N	0.8584681
CAF	Central African Republic	N	0.8071204
CUB	Cuba	N	0.8908176
CPV	Cape Verde	N	0.6142804
CYP	Cyprus	N	0.5468579

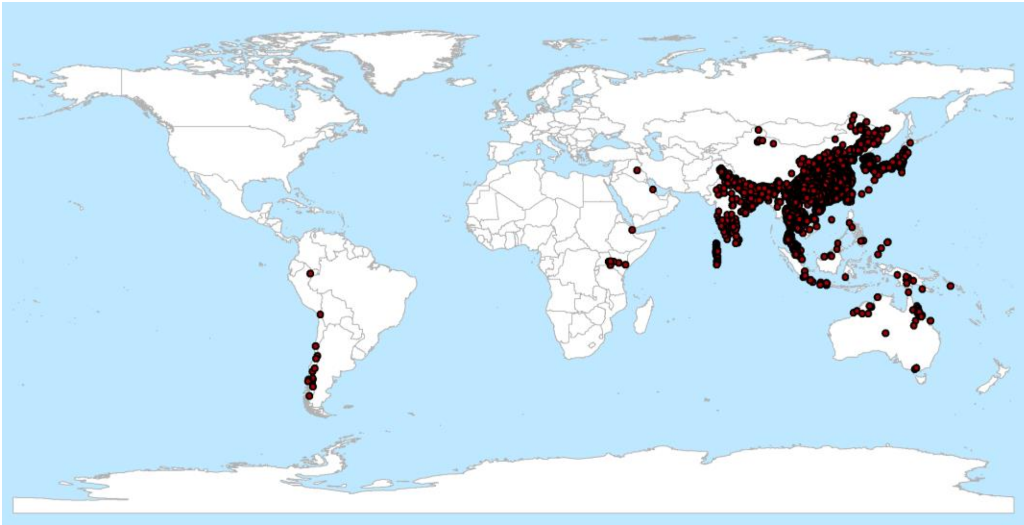
DMA	Dominica	N	0.8221209
DOM	Dominican Republic	N	0.7919364
ECU	Ecuador	N	0.6969346
EGY	Egypt	N	0.6297006
GNQ	Equatorial Guinea	N	0.9001058
ERI	Eritrea	N	0.6770846
SLV	El Salvador	N	0.9009198
ETH	Ethiopia	N	0.6561315
GUF	French Guiana	N	0.9039915
FJI	Fiji	N	0.7252437
FSM	Federated States of Micronesia	N	0.6155528
PYF	French Polynesia	N	0.5984889
GMB	Gambia, The	N	0.8105883
GAB	Gabon	N	0.9007281
GHA	Ghana	N	0.8620615
GRD	Grenada	N	0.7510616
GLP	Guadeloupe	N	0.8174323
GUM	Guam	N	0.7322897
GTM	Guatemala	N	0.8138361
GIN	Guinea	N	0.8109486
GUY	Guyana	N	0.8653698
ISR	Gaza Strip	N	0.6982115
HTI	Haiti	N	0.8759718
ISR	Israel	N	0.625336
CIV	Ivory Coast	N	0.8743957
IRQ	Iraq	N	0.626605
JAM	Jamaica	N	0.8452687
JOR	Jordan	N	0.5143677
KWT	Kuwait	N	0.6646774
LBN	Lebanon	N	0.5795968
LBR	Liberia	N	0.8993461
LUX	Luxembourg	N	0.5436459
LBY	Libya	N	0.5136242
MDG	Madagascar	N	0.6965172
MTQ	Martinique	N	0.7875484
MAC	Macau	N	0.9347954
MYT	Mayotte	N	0.6908741
MSR	Montserrat	N	0.8481452
MWI	Malawi	N	0.7359778
MLI	Mali	N	0.8172341
MUS	Mauritius	N	0.8245977
MRT	Mauritania	N	0.6924398
MLT	Malta	N	0.5239161
OMN	Oman	N	0.6924585
MEX	Mexico	N	0.6585243
MOZ	Mozambique	N	0.752112

NCL	New Caledonia	N	0.6026257
NER	Niger	N	0.7735726
NGA	Nigeria	N	0.888163
SUR	Suriname	N	0.9087343
NIC	Nicaragua	N	0.862294
PRY	Paraguay	N	0.7404584
PAN	Panama	N	0.8557939
GNB	Guinea-Bissau	N	0.6897403
QAT	Qatar	N	0.6716459
REU	Reunion	N	0.6210277
PRI	Puerto Rico	N	0.887071
RWA	Rwanda	N	0.6676883
SAU	Saudi Arabia	N	0.674943
KNA	St. Kitts and Nevis	N	0.533497
SYC	Seychelles	N	0.617013
ZAF	South Africa	N	0.5488808
SEN	Senegal	N	0.6978806
SVN	Slovenia	N	0.5009149
SLE	Sierra Leone	N	0.9076485
SGP	Singapore	N	0.7234231
SOM	Somalia	N	0.7661952
SDN	Sudan	N	0.7667642
SYR	Syria	N	0.5253849
ARE	United Arab Emirates	N	0.6762679
TTO	Trinidad and Tobago	N	0.8951436
TON	Tonga	N	0.6164699
TGO	Togo	N	0.8758947
TUN	Tunisia	N	0.5083447
TZA	Tanzania, United Republic of	N	0.7863342
UGA	Uganda	N	0.7717489
USA	United States	N	0.5070344
BFA	Burkina Faso	N	0.8282123
VCT	St. Vincent and the Grenadines	N	0.8325873
VEN	Venezuela	N	0.8323584
VIR	Virgin Islands	N	0.6420539
NAM	Namibia	N	0.5602835
ISR	West Bank	N	0.5082296
ESH	Western Sahara	N	0.543527
WSM	Western Samoa	N	0.8119532
SWZ	Swaziland	N	0.695571
YEM	Yemen	N	0.6732206
ZMB	Zambia	N	0.6338152
ZWE	Zimbabwe	N	0.6090861

Table S11: Predicted high environmental suitability areas and populations living in the environmental suitable areas in 2001, 2010 and 2020.

Worldwide				
	Area (km^2)	Uncertainty	Change	
2001	66,373,650	5361,0975-82,853,000	2001-2010	716,975
2010	67,090,625	54,485,500-82,875,850	2010-2020	873,350
2020	67,963,975	55,315,825-85,155,000	2001-2020	1590,325
	Population	Uncertainty	Change	
2001	3,163,442,997	2,749,422,428-3,557,279,550	2001-2010	558,968,110
2010	3,722,411,107	3,284,405,622-4,137,388,957	2010-2020	679,578,866
2020	4,401,989,973	3,855,816,571-4,895,784,723	2001-2020	1,238,546,976
Countries with confirmed human case				
	Area (km^2)	Uncertainty	Change	
2001	21,356,975	17,483,675-27,371,200	2001-2010	-309,00
2010	21,326,075	17,789,200-27,182,650	2010-2020	805,725
2020	22,131,800	18,477,275-28,172,625	2001-2020	774,825
	Population	Uncertainty	Change	
2001	2,039,459,989	1,915,091,037-2,145,480,086	2001-2010	312,157,042
2010	2,351,617,031	2,221,678,479-2,462,458,144	2010-2020	192,130,634
2020	2,543,747,665	2,428,123,320-2,687,674,384	2001-2020	504,287,676

Supplementary Figures



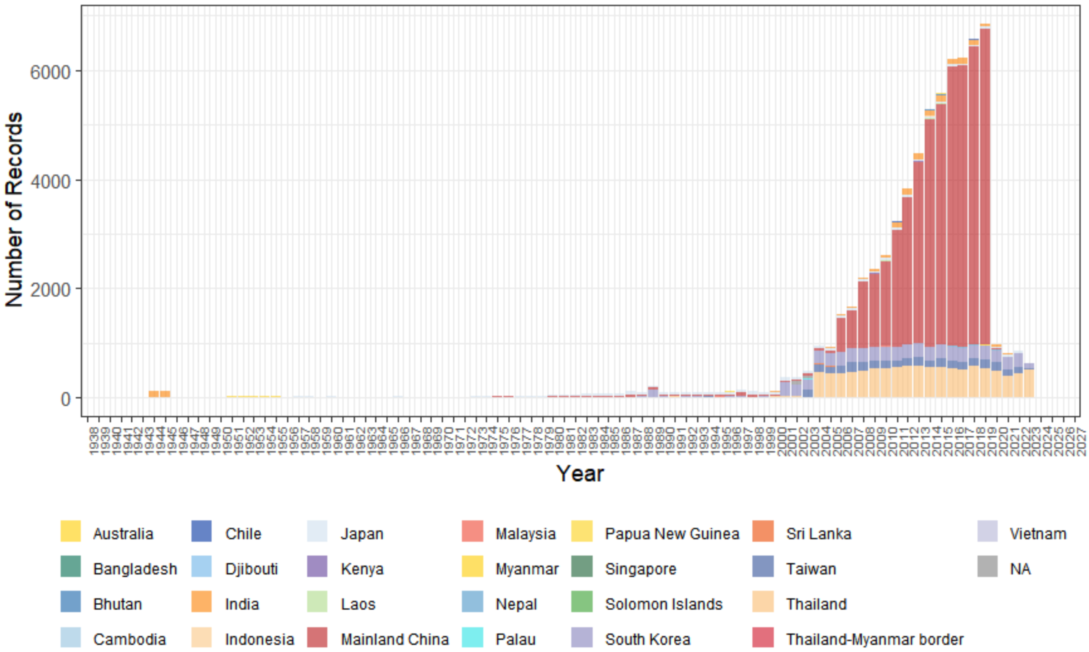


Figure S1 (a) Locations of the occurrence records used in model development, (b) number of unique scrub typhus occurrence locations per year by country/region.

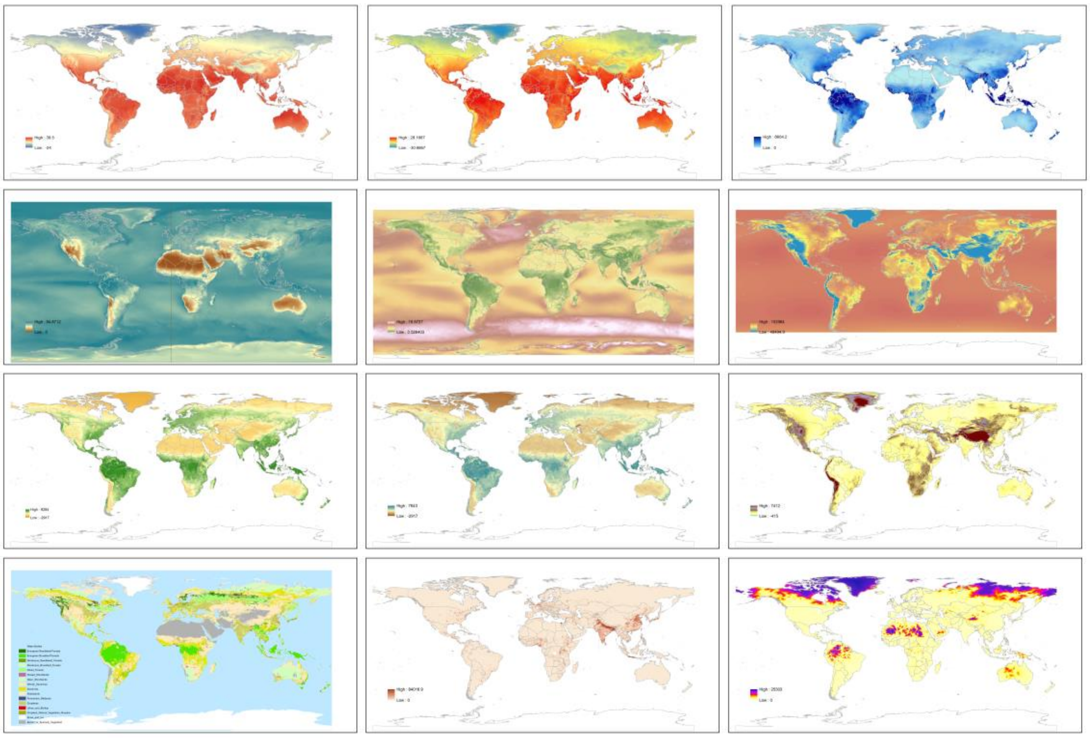


Figure S2 Covariates included in the ensemble models (2020). The covariates are listed in order from left to right and top to bottom as follows: max temperature, min temperature, precipitation, relative humidity, wind speed, air pressure, NDVI, EVI, elevation, 17 types of land use, population density, and urbanization (measured as travel time to major cities).

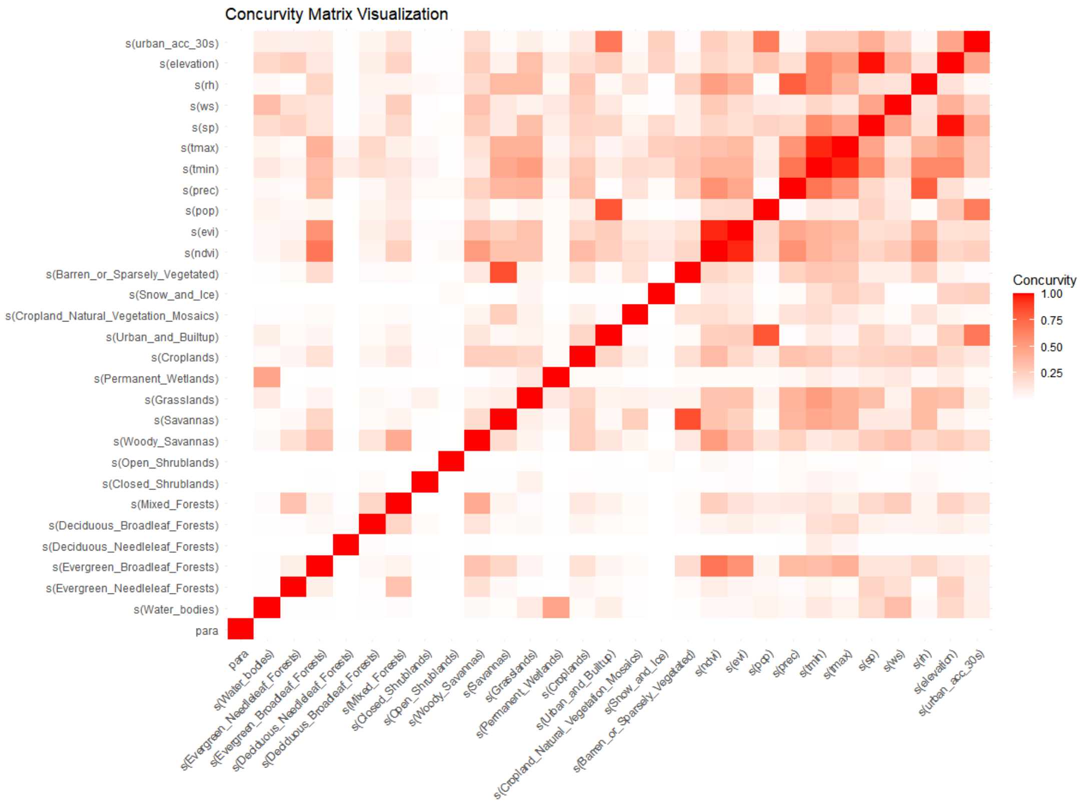


Figure S3 Concurvity analysis result for GAM

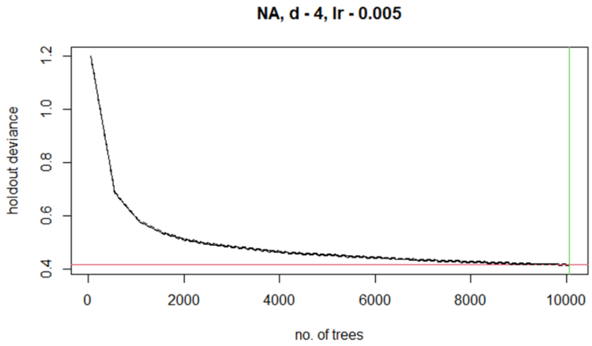


Figure S4 BRT parameter tuning result

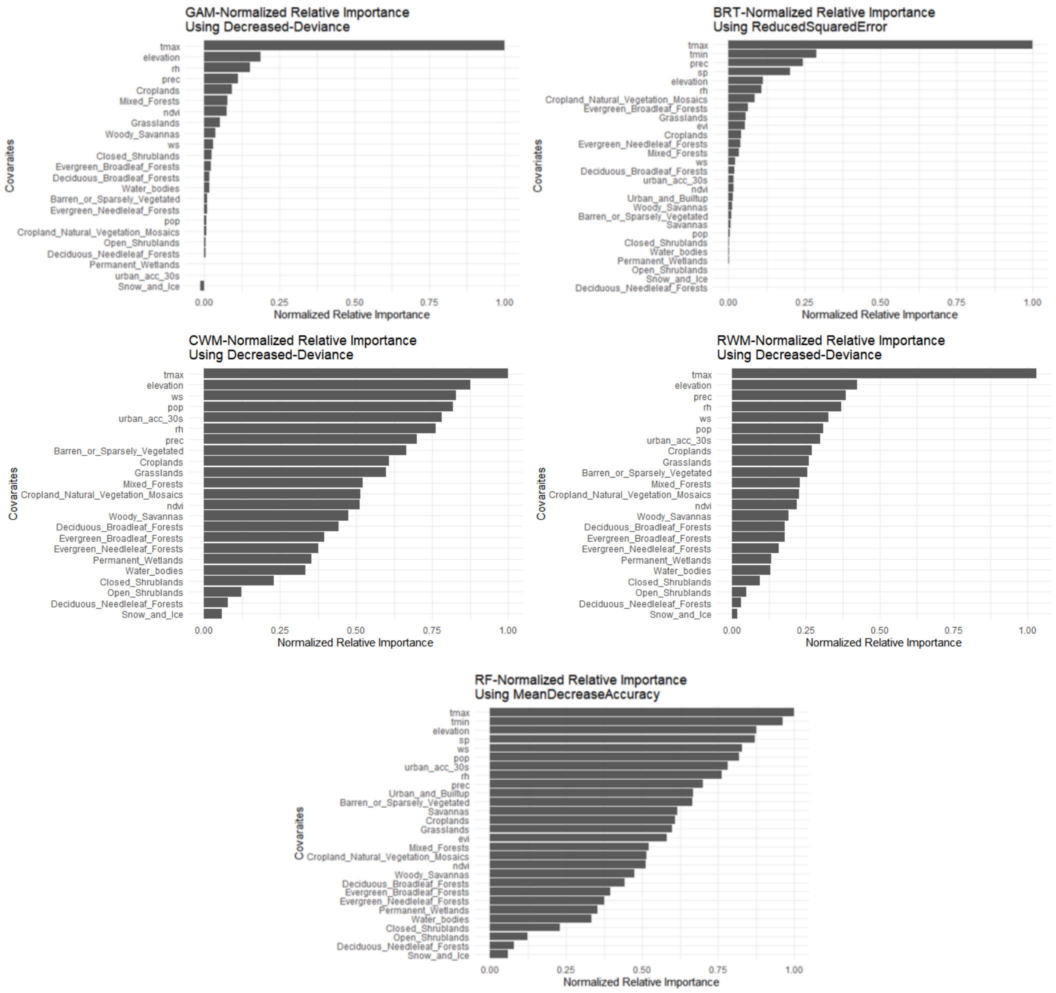
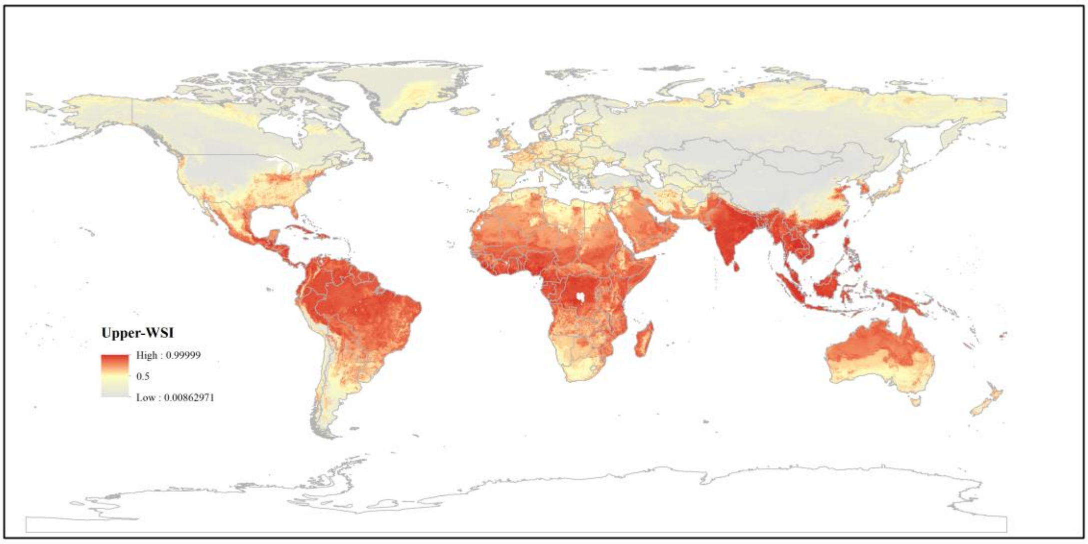
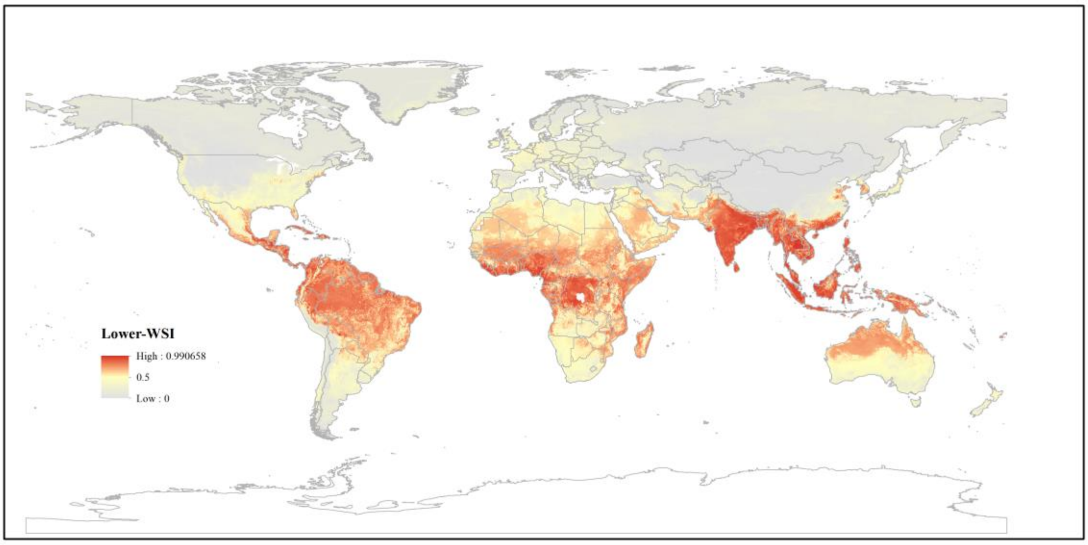


Figure S5 Relative importance of covariates





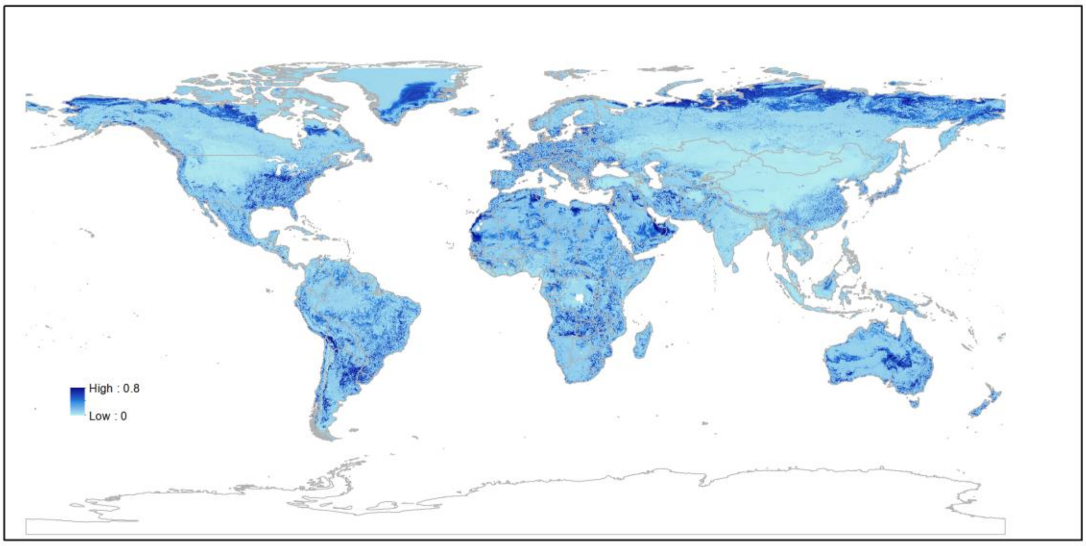


Figure S6 Maps of uncertainty in environmental suitability estimates (95%CI upper, 95%CI lower and width of the confidence interval).

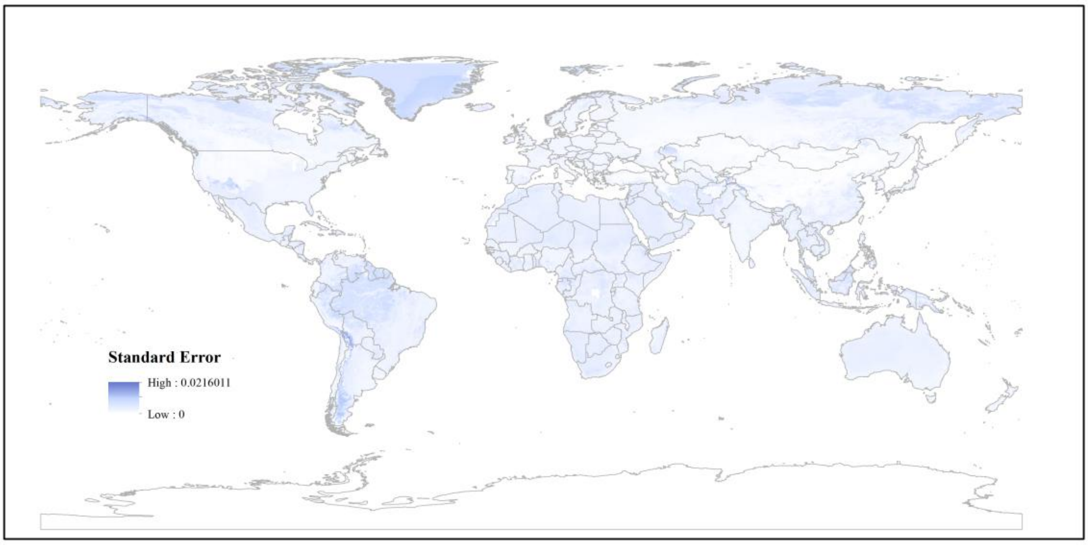


Figure S7 Standard error of robustness test result 100 bootstrap sampling.

References

1. Li Z, Xin H, Sun J, et al. Epidemiologic Changes of Scrub Typhus in China, 1952-2016. *Emerging infectious diseases* 2020; **26**(6): 1091-101.

2. Yoshikura H. Seasonality and Geographical Distribution of Tsutsugamushi Diseases in Japan: Analysis of the Trends since 1955 till 2014. *Japanese journal of infectious diseases* 2018; **71**(1): 1-7.

3. Kinoshita H, Arima Y, Shigematsu M, et al. Descriptive epidemiology of rickettsial infections in Japan: Scrub typhus and Japanese spotted fever, 2007-2016. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2021; **105**: 560-6.

4. Park S, Cho E. National Infectious Diseases Surveillance data of South Korea. *Epidemiol Health* 2014; **36**: e2014030.

5. Lin FH, Chou YC, Chien WC, Chung CH, Hsieh CJ, Yu CP. Epidemiology and risk factors for notifiable scrub typhus in Taiwan during the period 2010–2019. *Healthcare(Switzerland)* 2021; **9**(12).

6. Wangrangsimakul T, Elliott I, Nedsuwan S, et al. The estimated burden of scrub typhus in Thailand from national surveillance data(2003-2018). *PLoS neglected tropical diseases* 2020; **14**(4): e0008233.

7. Weiss DJ, Nelson A, Gibson HS, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* 2018; **553**(7688): 333-6.