Relative Importance of Environmental and Spatial Predictors for Slipper Orchid (*Paphiopedilum* spp.) Distribution in Thailand Using Random Forest

Pholgerddee, P.,[1,2] Pattanakiat, S.,[1] Nakmuenwai, P.,[1] Sattraburut, T.[1] and Phutthai, T.[1*]
[1]Faculty of Environment and Resource Studies, Mahidol University,
Nakhon Pathom, Thailand, E-mail: sura.pat@mahidol.ac.th, pisut.nak@mahidol.ac.th,
thunyapat.sat@mahidol.ac.th, thamarat.phu@mahidol.ac.th*
[2]Department of Geoinformatics, Faculty of Humanities and Social Sciences, Burapha University, Chon Buri, Thailand, E-mail: pichitporn@go.buu.ac.th
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v22i4.4945

Abstract
Slipper orchids (Paphiopedilum spp.) represent some of Southeast Asia's most ecologically specialized and conservation-sensitive orchid taxa. Their distribution in Thailand is shaped by intricate environmental gradients, encompassing climate, topography, vegetation structure, and geographic location. Nonetheless, comprehensive national assessments of habitat appropriateness for Paphiopedilum species are scarce. This research utilized a species distribution modelling methodology to ascertain critical environmental factors and forecast appropriate habitats for Paphiopedilum species throughout Thailand. Seventy occurrence reports were obtained from biodiversity databases, field surveys, and herbarium collections. A Random Forest modelling framework was employed utilizing twelve environmental predictors that encompass climatic, non-climatic, and spatial factors. A balanced presence-absence dataset was created, and model performance was assessed using various statistical criteria, including the Brier score. The model attained a Brier score of 0.05, signifying elevated predictive accuracy. Climatic factors accounted for 38.30% of overall variable importance, followed by non-climatic variables at 37.40% and spatial predictors at 24.30%. The model exhibited robust predictive efficacy, with an AUC of 0.94, sensitivity of 0.80, specificity of 0.76, and an overall classification accuracy of 0.78. Principal predictors encompassed latitude, longitude, elevation, vegetation structure (NDVI), and temperature-associated factors. Optimal habitats were predominantly situated in the mountainous areas of northern and western Thailand, with supplementary fragmented patches observed in certain locations of southern Thailand. The results underscore the significant influence of topographic gradients, temperature, and vegetation structure on the ecological niche of Paphiopedilum species. These findings establish a scientific foundation for pinpointing priority conservation zones and facilitating long-term management strategies for slipper orchid ecosystems in Thailand.

Keywords: Environmental Predictors, Geospatial Analysis, Random Forest, Slipper Orchids, Species Distribution

1. Introduction

This study focuses on the genus *Paphiopedilum*, commonly known as slipper orchids, which represents a particularly distinct group within the orchid family (*Orchidaceae*), both with its characteristic flower morphology and horticultural value. The Orchidaceae family is one of the biggest groups of flowering plants, containing over 28,000 species worldwide. It also displays incredible biological diversity in tropical and subtropical environments [1] and [2]. As a result of its diverse landscape, variable conditions and forest types, the Southeast Asia region has been defined as an important hotspot for orchid diversity [3]. In particular, *Paphiopedilum* species are characterized by ecological specialization and a need for restricted habitats [4].

Paphiopedilum orchids tend to live in shaded forest understories, limestone hills, and montane environments, where temperature, humidity, soil conditions, and vegetation structure are beneficial for their cultivation or reproduction [4] and [5]. Such orchids often possess narrow ecological tolerances

and limited dispersal abilities, heightening their vulnerability to environmental changes [5]. Previous ecological studies have shown that temperature, rainfall, elevation, vegetation cover and so on are important drivers of orchid distribution [6]. The most prominent are mountain habitats, which provide relatively stable microspatial conditions with cooler temperatures and higher humidity as well as lower thermal extremes favored by slipper orchid survival [5] and [6]. Southeast Asia represents one of the major global centers of orchid diversity, supporting many species across a wide range of tropical forest ecosystems, with Thailand as a central location of it. There is such a variety of orchid species that thrive in various conditions: tropical evergreen forests, mixed deciduous forests and "montane" [3]. In Thailand, several *Paphiopedilum* species have been reported, many of which are categorized as rare or endangered due to habitat destruction from logging activities and illegal collection for the ornamental plant trade [4] and [7]. The loss of suitable habitats for many orchids has been accelerated dramatically by agriculture, infrastructure development and forest disturbance. Slipper orchids are particularly sensitive to environmental and anthropogenic disturbances because they are slow-growing and have particular habitat requirements [5] and [6].

Therefore, it is crucial to understand what environmental factors determine the distributions of orchids in order to conserve biodiversity and manage habitats. Species Distribution Modelling (SDM) become a strong method for investigating species environment relationships and predicting habitat suitability in recent decades [8] and [9]. SDMs integrate species presence and absence records with environmental factors with environmental factors to identify where these habitats are considered optimal for recreation and what biophysical variables affect their distribution [10]. These models are widely used in ecology and conservation planning because they provide quantitative information about the spatial distribution of biodiversity, as well as the environmental drivers that affect species distribution [7] and [11]. Out of all the modelling techniques applied to SDM, machine learning algorithms have received much attention lately, as they have demonstrated a capacity to describe complex non-linear relationships between environmental variables [12]. Random Forest (RF) is one of the most popular machine learning methods for ecological modelling because it can both make accurate predictions and help figure out which predictor variables are most important [13]. RF creates an ensemble of decision trees trained on data generated via bootstrap sampling, and by selecting a random variable subset to search for the best split [13]. Due to these benefits, Random Forest has been widely used in various ecological studies to analyze species distribution and detect significant environmental predictors [14] and [15].

Despite the great advent of species distribution modeling for many plant taxa, studies dedicated exclusively to the distribution of *Paphiopedilum* species are still few and far between, especially when we refer to national spatial scales in Southeast Asia. Prior studies have largely assessed environmental and climate factors in isolation or are specific to localized geographic areas, while most research examining how climatic, non-climatic, and spatial variables interact has been confined to smaller spatial extents. Assessments of the relative importance of predictors provide vital information for interpreting and better linking species–environment relationships to specific ecological drivers structuring orchid habitats [7] and [16].

The objective of this investigation was employing a RF modelling framework to look at how climatic, non-climatic, and geographical factors affect the distribution of slipper orchid species in Thailand. Using multiple environmental datasets that consist of climatic variables, topographic characteristics, vegetation indices, soil properties and geological coordinates to identify habitat optimum by key environmental gradients. These results are valuable data toward gaining a better understanding of the environmental drivers of slipper orchid distribution and towards improving spatial conservation planning for these ecologically and horticulturally important species.

2. Methodology

The methodology was designed to ensure reproducibility and follow established standards in species distribution modeling, geospatial preprocessing, and machine-learning analysis. All steps were performed using RStudio with fully documented workflows and fixed random seeds to guarantee consistency across analyses.

2.1 Study Area

The study encompassed the entirety of Thailand, which spans diverse physiographic and climatic gradients from mountainous uplands in the north and west to lowland plains and peninsular landscapes in the south [17] [18] and [19]. The spatial extent of study area and distribution of *Paphiopedilum* occurrence records are illustrated in Figure 1. This environmental heterogeneity provides an appropriate geographic context for evaluating broad-scale habitat associations of *Paphiopedilum* spp.

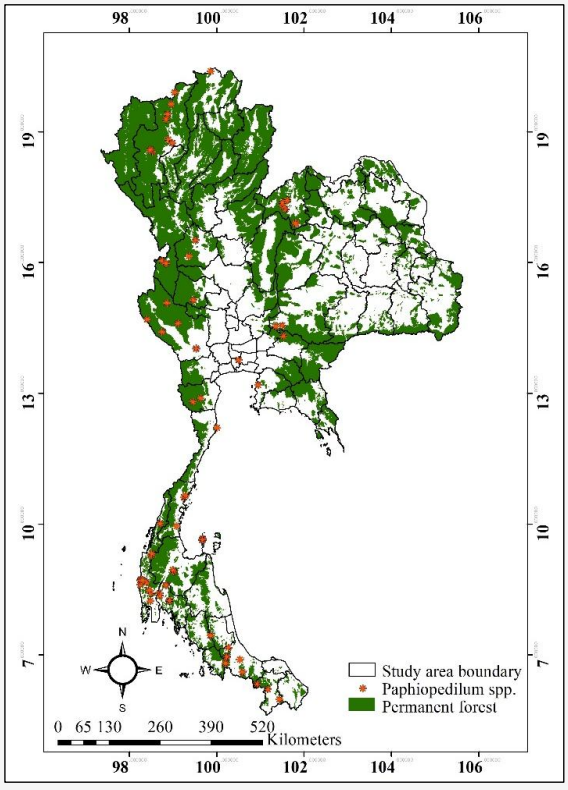

Figure 1: Study area showing the distribution of *Paphiopedilum* species occurrences and permanent forest areas across Thailand

At the national scale, the distribution of slipper orchids is expected to be influenced by variation in temperature regime, moisture availability, topography, vegetation structure, and spatial location [7] and [20]. A national boundary shapefile was used to define the analysis extent and to crop and mask all environmental layers to the study area, ensuring spatial consistency across all predictor datasets used in the modelling framework [21].

2.2 Species Occurrence Data

Occurrence records of *Paphiopedilum* spp. in Thailand were compiled from different validated botanical data sources, including herbarium collections, global biodiversity databases, and observations from the field. Geographic coordinates for approximately 14 *Paphiopedilum* species were extracted from several international and regional herbariums, including the Aarhus University Herbarium (AAU, Denmark), Andalas University Herbarium (ABD, Indonesia), The Natural History Museum Herbarium (BM, United Kingdom), Chulalongkorn University Herbarium (BK, Thailand), Forest Herbarium of the Department of National Parks, Wildlife and Plant Conservation (BKF, Thailand), Royal Botanic Garden Edinburgh Herbarium (E, United Kingdom), Royal Botanic Gardens Kew Herbarium (K, United Kingdom), Herbarium of the Kunming Institute of Botany (KUN, China), Herbarium of Lund University (L, Sweden), Muséum National d'Histoire Naturelle Herbarium (P, France), Herbarium of the University of Puerto Rico (PE, Puerto Rico), Prince of Songkla University Herbarium (PSU, Thailand), Queen Sirikit Botanic Garden Herbarium (QBG, Thailand), and Singapore Botanic Gardens Herbarium (SING, Singapore).

Additional occurrence records were obtained from the Global Biodiversity Information Facility (GBIF) database and relevant published sources to supplement herbarium data and improve spatial coverage of species records [22]. Field survey data were also incorporated to enhance the dataset and improve representation of known species localities within Thailand. All occurrence records were compiled in comma-separated value (CSV) format with geographic coordinates referenced to the WGS84 coordinate system. Data cleaning procedures were performed to improve positional reliability and reduce sampling bias. Records with missing or invalid geographic coordinates, records located outside the study area, and repeated occurrences that were contained within the same raster grid cell were eliminated. In addition, recordings associated with cultivated plants, botanical gardens, or clearly unrealistic geographic locations were identified and excluded from the dataset. These filtering procedures followed commonly recommended practices in species distribution modelling to reduce georeferencing errors and spatial sampling artefacts [23] and [24].

We retained and used 70 validated occurrence records as presence data for the species distribution modelling analysis following data cleaning. We then mapped the spatial distribution of these occurrence points and extracted environmental predictor values for modelling. However, while the number of occurrence records is not particularly large, herbarium-based datasets are commonly employed in spatial modelling studies for rare or region-restricted plant species (especially where no field observations are available [25].

2.3 Environmental Predictors

The overall methodological workflow of this study is illustrated in Figure 2. The workflow consists of five main steps: (1) compilation of species occurrence data, (2) acquisition and preprocessing of environmental predictor variables, (3) multicollinearity analysis and variable selection, (4) Random Forest model development and tuning, and (5) model evaluation and interpretation of variable

importance. This structured workflow ensures reproducibility and consistency across all stages of the analysis and provides a clear framework for understanding the modelling process. This research employs two categories of data: records of species occurrences and environmental predictor variables. Environmental predictors were grouped into three categories: (1) climatic variables, (2) non-climatic environmental variables and (3) spatial variables. All environmental predictors used in the modelling process are summarized in Table 1.

Each environmental predictor was chosen for its ecological significance to orchid habitat needs and its recognized application in species distribution modelling research. Temperature-related variables, such as annual mean temperature (BIO1) and maximum temperature of the warmest month (BIO5), are essential factors affecting plant physiological processes, hence influencing the growth, flowering, and survival of orchid species in tropical ecosystems [8] and [23]. Precipitation variables, including annual precipitation (BIO12) and precipitation during the driest quarter (BIO17), indicate moisture availability, which is crucial for orchid development, particularly for species residing in humid forest habitats [11]. Topographic factors, such as elevation (DEM) and slope, are linked to microclimatic variation, water drainage, and habitat diversity.

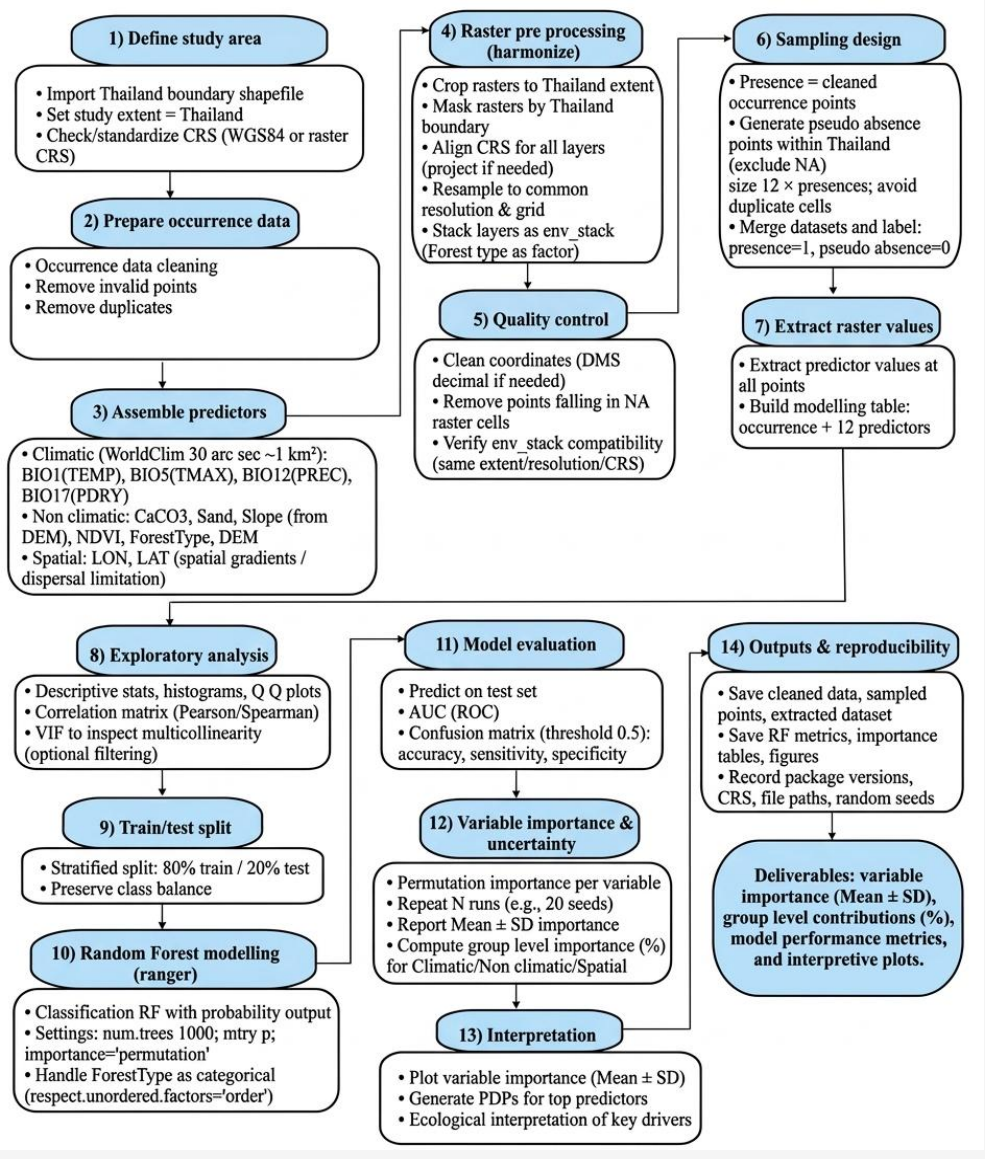

Figure 2: Environmental and spatial predictors for slipper orchid (*Paphiopedilum spp.)* distribution study workflow

Table 1: Environmental predictors used in the Random Forest model

Variable	Code	Category	Data source	Year	Value range	Reference
Annual mean temperature	BIO1	Climatic	WorldClim global climate database	1970 to 2000	15.2 to 28.9 °C	[26]
Maximum temperature of warmest month	BIO5	Climatic	WorldClim global climate database	1970 to 2000	24 to 38.4 °C	[26]
Annual precipitation	BIO12	Climatic	WorldClim global climate database	1970 to 2000	774 to 4,553 mm	[26]
Precipitation of driest quarter	BIO17	Climatic	WorldClim global climate database	1970 to 2000	8 to 460 mm	[26]
Normalized Difference Vegetation Index	NDVI	Vegetation	MODIS satellite NDVI product (MOD13A2)	2023 (Oct–Nov)	0.3 to 0.9	[26]
Forest type	FT	Land cover	Royal Forest Department (RFD), Thailand	2019	Categorical classes	[17]
Calcium carbonate	$CaCO_3$	Soil	Land Development Department soil database (Thailand)	2019	0 to 8.5%	–
Sand fraction	SAND	Soil	Land Development Department soil database (Thailand)	2019	0 to 39.5%	–
Slope	SLOPE	Topographic	Hydro1K GTOPO30 DEM (EROS)	1996	0 to 3,211 m	[23]
Elevation	DEM	Topographic	Hydro1K GTOPO30 DEM (EROS)	1996	0 – 2,800 m	[23]
Latitude	LAT	Spatial	Geographic coordinate	–	–	–
Longitude	LON	Spatial	Geographic coordinate	–	–	–

Elevation gradients are crucial for montane orchids, which often inhabit cooler and more stable climatic environments [5] and [6]. The vegetation structure, shown by NDVI, acts as a surrogate for canopy density and productivity, affecting light availability and microhabitat conditions for understory orchid species [27]. Soil variables, such as calcium carbonate ($CaCO_3$) and sand fraction, were used to depict edaphic conditions. Numerous *Paphiopedilum* species are linked to limestone substrates, where calcium availability is essential for plant establishment and nutrient dynamics. Land-cover data regarding forest types offer insights into habitat structure and ecosystem composition, which are crucial for comprehending species–habitat relationships at the landscape level. Spatial variables (latitude and longitude) were incorporated to elucidate extensive geographic gradients and spatial structures affecting species distribution, as previously mentioned as shown in Figure 3.

The climatic variables were extracted from the WorldClim global climate dataset with a spatial resolution of 30 arc-seconds (~1 km.) [28]. Four bioclimatic variables were chosen due to their ecological relevance to plant distributions and relatively low multicollinearity: annual mean temperature (BIO1), maximum temperature of the warmest month (BIO5), annual precipitation (BIO12), and precipitation of the driest quarter (BIO17). These variables are important thermal and moisture conditions that impact plant physiological processes, species occurrence/outcome, and habitat suitability [6] and [7], and they are commonly selected in modelling species' distribution studies. To evaluate local habitat characteristics that might affect the distribution of *Paphiopedilum* species, there are received non-climatic environmental variables. Topographic variables (e.g., elevation and slope), calculated from a digital elevation model (DEM), represent terrain structure, which is closely related to local microclimatic conditions, including temperature variation, humidity and water drainage. Elevation gradients are especially relevant to tropical mountainous regions in which many orchid species are found, often inhabiting cool and moist upland environments. Topographic characteristics, such as elevation and slope, were obtained from the Hydro1K GTOPO30 digital elevation model (DEM) dataset, which has an original resolution of about 1 km. The datasets were resampled and aligned to correspond with the geographical resolution of other environmental predictors utilized in the modelling framework. The Normalized Difference Vegetation Index (NDVI) was obtained from the MODIS MOD13A2 product, which has a spatial resolution of 1 km. The NDVI data utilized in this study were obtained from 1 October 2023 to 30 November 2023, aligning with the post-monsoon period in Thailand.

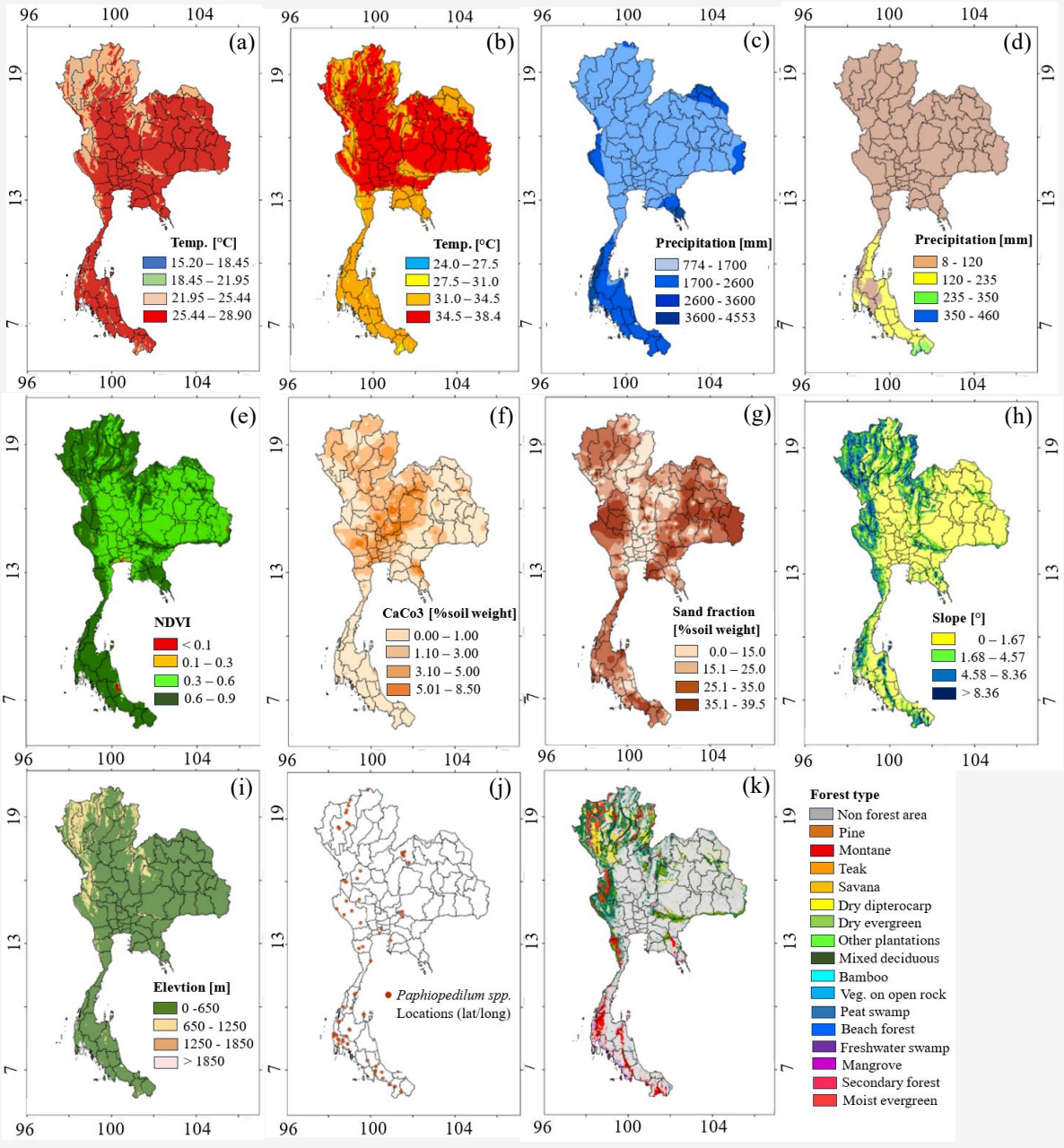

Figure 3: Thematic maps of influencing factors (a) annual mean temperature, (b) maximum temperature of warmest month, (c) annual precipitation, (d) precipitation of driest quarter, (e) NDVI, (f) $CaCo_3$, (g) sand, (h) slope, (i) elevation, (j) latitude/longitude, and (k) forest type

This timeframe aligns with the zenith of vegetative productivity and optimal canopy formation in tropical forests, therefore offering an accurate depiction of forest structure and habitat circumstances pertinent to understory orchid species [15] and [27]. Despite NDVI's seasonal fluctuations, the chosen timeframe reflects optimal vegetation conditions and offers a uniform representation of habitat structure. Consequently, NDVI serves as a surrogate for vegetative attributes rather than a temporally variable prediction. To ensure consistency with other environmental predictors (e.g., WorldClim variables at ~1 km resolution), all raster datasets were resampled and aligned to a common spatial resolution and grid system. This preprocessing step minimizes potential bias arising from spatial resolution mismatch among datasets.

Soil variables were also incorporated as proxies of edaphic conditions that could affect orchid distribution. Soil calcium carbonate ($CaCO_3$) and soil

sand fractions were especially included in the analysis. Multiple species of *Paphiopedilum* are documented from limestone-derived habitats, and calcium-rich substrates are important in plant establishment and nutrient availability. Including $CaCO_3$ thus allows the model to capture substrate-related environmental conditions that may affect slipper orchid habitat suitability. National land-cover datasets were used as an indication of major vegetation classes across Thailand; in the absence of available forest-type data, they were obtained from national land-cover datasets. These forest categories offer extra knowledge on habitat structure and landscape context, which potentially affect patterns of orchid occurrence. In addition to environmental predictors, spatial variables (latitude and longitude) were used as proxies to represent extensive geographic gradients and spatial organization. These patterns may not be entirely elucidated by environmental variables alone. These factors may elucidate regional autocorrelation, dispersion limitations, and unquantified environmental variations affecting species distribution [6] and [14].

Before modelling, all environmental raster layers were resampled and aligned to a consistent geographic resolution and coordinate reference systems. This preprocessing step also provided spatial consistency for predictor variables and allowed reliable extraction of environmental values at species occurrence locations used in the modelling framework [8].

2.4 Multicollinearity Analysis

Before model development, multicollinearity amongst environmental predictors was assessed to decrease redundancy and increase interpretability of the models. That high correlation of predictors can cause bias in model estimation and inflate the perceived importance of environmental variables in modelling species distributions [29]. To identify strongly correlated variable pairs, pairwise Pearson correlation coefficients were calculated between all continuous predictor variables. If correlation coefficients were above a threshold of $|r| \geq 0.7$, variables were inspected closely, and one of the correlated variables was disincluded based on ecological relevance and interpretability of the data. Furthermore, the Variance Inflation Factor (VIF) was calculated to further check for multicollinearity between predictors.

In general, VIF values of greater than 10 demonstrate problematic multicollinearity, which means the variables with such high correlation were removed from further analysis. This process ensured that the final set of environmental predictors included in the RF model did not exhibit strong correlation and each variable added independent information to the modelling approach [29] and [30]. The environmental variables that remained after multicollinearity screening were retained for species distribution modelling, followed by a variable importance analysis. The VIF was utilized to further assess multicollinearity among environmental predictors. All remaining predictors demonstrated VIF values beneath the widely recognized threshold of 10, signifying the absence of severe multicollinearity in the final dataset. In addition, pairwise Pearson correlation coefficients among predictor variables were generally below $|r| = 0.7$, suggesting that each environmental variable contributed independent information to the modelling framework. Therefore, all selected predictors were retained for the Random Forest modelling analysis. The detailed VIF values of the environmental predictors are presented in Table 2.

Table 2: Variance Inflation Factor (VIF) values for environmental predictors used in the Random Forest model

Variable	VIF
LAT	2.34
LON	2.28
DEM	2.67
Slope	1.94
TEMP	3.21
TMAX	3.08
PREC	2.16
PDRY	2.44
NDVI	2.02
$CaCO_3$	1.71
Sand	1.63
Forest type	1.58

2.5 Pseudo-absence Sampling

Random Forest classification needs presence and absence data were collected, resulting in the generation of pseudo-absence points throughout the study area of study to represent environmental conditions where it is assumed that species are absent. When true absence data cannot be easily obtained for rare species, pseudo-absence generation is a common and widely used approach in species distribution modelling [6] [30] and [31]. After removing locations where there was missing environmental data and cells that corresponded to presence records, pseudo-absence points were drawn randomly from environmentally valid raster cells within the study area. The quantity of pseudo-absences was determined by need to achieve a balanced ratio between presence and absence classes in the modelling dataset. For construction of the classification dataset used in model training and

tests, 70 presence records were combined with randomly generated pseudo-absence points.

By randomly sampling the environmental space of the study region, the model compares environmental conditions correlated with species occurrences to a wider context. It has been suggested in species distribution modelling [30] and [32] that using background points to reduce sampling bias can lead to better stable model summary parameters when true absence data are not available. In this study, a total of 70 validated occurrence records were used as presence data. To construct a balanced classification dataset for the Random Forest model, 1,000 pseudo-absence points were randomly generated across the study area after excluding cells containing presence records and areas with missing environmental data. The selection of a larger number of pseudo absence points relative to presence data is commonly recommended in species distribution modelling studies to improve model stability and represent the environmental background more effectively [30] and [33]. Random sampling across the environmental space of Thailand allows the model to contrast conditions associated with species occurrences with the broader environmental gradients present in the study region.

2.6 Random Forest Modelling

Species distribution modelling was conducted with the Random Forest (RF) approach using the R statistical computing environment. RF is a machine learning ensemble technique that employs bootstrapped samples to construct numerous classification trees and random subsets of predictor variables [32]. It combines the results from all trees in the ensemble to give a final prediction, enhancing predictive performance and mitigating overfitting. Random Forest has found extensive applications in ecological modelling, given its ability to accommodate nonlinear relationships, complex interplay between predictors and mixed types of variables without the need for strict parametric assumptions [33]. The algorithm also returns estimates of the importance of individual variables, which can inform researchers about how much weight they give to their environmental predictors in their model performance as shown in Table 3. RF model in this study was trained using climate, non-climatic and spatial predictors. Permutation importance was used to evaluate variable importance, measuring changes in model accuracy when predictor values are permuted randomly [32] and [34].

To improve model robustness, a limited hyperparameter tuning framework was applied. Candidate values of key Random Forest parameters are presented in Table 4. The final model configuration was selected based on overall model performance and stability. The selected parameter values are summarized in Table 5. A limited hyperparameter tuning process was done to make the model more stable and avoid overfitting. This was done by testing different combinations of important Random Forest parameters, such as the number of trees (ntree), the number of variables randomly chosen at each split (mtry), and the minimum node size. Table 4 shows the possible values for the candidate parameters. We chose the best combination of parameters based on how well the model worked (AUC) and how stable it was throughout multiple runs. Table 5 shows the final chosen settings for the parameters. The chosen settings (ntree = 1000, mtry = 3, min.node.size = 5) struck a decent mix between how well the model could forecast and how well it could generalize, so it wasn't too simple or too complex.

Table 3: Random Forest model parameters used in this

Parameter	Value	Description
Number of trees (ntree)	1000	Number of decision trees in the Random Forest model
Number of variables per split (mtry)	3	Number of predictors randomly selected at each split
Importance measure	Permutation importance	Used to assess variable importance
Sampling strategy	Bootstrap sampling	Random sampling with replacement
Data split	80% training / 20% testing	Used for model evaluation

Table 4: Candidate hyperparameter settings used for Random Forest tuning

Parameter	Candidate values tested	Description
Number of trees (num.trees)	500, 750, 1000	Number of predictors randomly selected at each split
Variables per split (mtry)	2, 3, 4	Number of predictors randomly selected at each split
Minimum node size (min.node.size)	1, 5, 10	Minimum number of samples in terminal nodes
Split rule	Gini	Splitting criterion for classification trees
Importance measure	Permutation	Variable importance estimation method

Table 5: Final Random Forest settings selected for the study

Parameter	Selected value
num.trees	1,000
mtry	3
min.node.size	5
Splitrule	Gini
Importance	Permutation

2.7 Model Evaluation

Various prevalent classification and discrimination metrics were utilised to assess model performance, including the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, total accuracy, and Brier score. A confusion matrix was created to evaluate categorization performance according to a specified probability threshold. The AUC statistics assesses the model's capacity to differentiate between presence and absence locations, whilst sensitivity and specificity quantify the accuracy of forecasting presences and absences, respectively. The Brier score evaluates the precision of probabilistic forecasts by measuring the disparity between projected probabilities and actual results. An optimal classification threshold was determined using the maximum True Skill Statistic (TSS), which identifies the threshold that maximizes the balance between sensitivity and specificity [29]. This data-driven approach avoids the use of arbitrary thresholds (e.g., 0.5) and is widely recommended in species distribution modelling studies. It is important to note that threshold-independent metrics such as AUC were used to evaluate model discrimination ability, while threshold-dependent metrics (e.g., sensitivity, specificity, TSS, and accuracy) were calculated based on the TSS-optimized threshold. Model performance was evaluated using many metrics, hence diminishing dependence on any singular performance indication [35] and [36]. The TSS-based threshold selection is especially appropriate for ecological applications, as it considers both omission and commission mistakes.

2.8 Software Implementation

All spatial data processing and modelling steps were performed in the R statistical computing environment. The bundle workflow and R serve as flexible platforms for spatial data analysis, ecological modelling, and reproducible research workflows [38]. This involved basic geospatial data processing, including raster manipulation, cropping and masking, as well as extraction of environmental variables using widely used spatial analysis packages (e.g., terra, raster, sf). These packages allow for large spatial datasets to be handled efficiently, as well as vector and raster data structures that can be integrated into ecological modelling workflows.

Implementation of species distribution modelling using the Random Forest algorithm utilizes machine learning libraries and packages available in the R environment that allow model training, prediction, and evaluation. Statistical modelling functions, as well as methods suited for classification analysis, were used to carry out these tasks. The modelling workflow comprised extraction of environmental variables for species occurrence sites, construction of presence–pseudo-absence datasets, fitting of Random Forest models and calculation of variable importance and model evaluation metrics [37]. All spatial layers were processed and analyzed within a single coordinate reference system, and modelling procedures were carried out using reproducible scripts to provide transparency and replicability in the analytical workflow [34].

3. Results and Discussion

3.1 Model Performance

The Random Forest (RF) model was applied to evaluate the relative importance of climatic, non-climatic, and spatial predictors influencing the distribution of *Paphiopedilum* spp. across Thailand. The model incorporated twelve environmental predictors representing climatic conditions, vegetation structure, soil properties, topography, and spatial gradients. The confusion matrix (Table 6) indicated that 51 presence records and 53 absence records were correctly classified, while 13 presence records and 17 absence records were misclassified. Model evaluation metrics indicated strong predictive efficacy. RF model produced a Brier score of 0.05, suggesting high calibration accuracy between predicted probabilities and observed species occurrences. Low Brier score values indicate that predicted suitability probabilities closely correspond to actual species presence and absence patterns, confirming that the model effectively distinguishes between suitable and unsuitable environmental conditions [35] and [36].

Table 6: Confusion matrix of the Random Forest model

	Observed Presence	Observed Absence
Predicted Presence	51	17
Predicted Absence	13	53

The strong predictive performance observed in this study is consistent with previous research demonstrating the effectiveness of Random Forest algorithms in ecological modelling, particularly for datasets characterized by nonlinear relationships and complex environmental interactions [6] and [13]. Machine-learning approaches such as RF are particularly well suited for species distribution modelling because they can integrate heterogeneous environmental predictors without strict parametric assumptions [33]. The assessment of model performance (Table 7) employed multiple complementary measures, including the Brier score, Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, and specificity. The Random Forest model achieved an AUC value of 0.94, signifying exceptional discrimination ability between presence and pseudo-absence sites. The ROC curve (Figure 4) reinforces the model's robust predictive efficacy. Employing the optimal threshold established through TSS maximization (0.47), the model attained a sensitivity of 0.80 and a specificity of 0.76. This yielded a TSS value of 0.56, signifying a balanced trade-off between omission and commission mistakes.

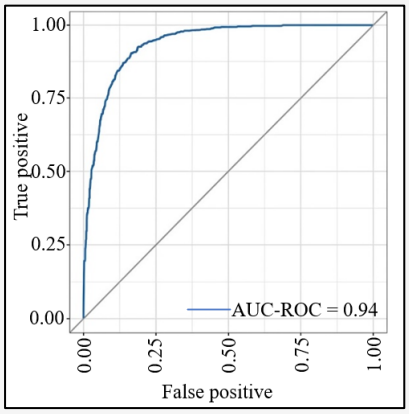

Figure 4: Receiver Operating Characteristic (ROC) curve of the Random Forest model for predicting *Paphiopedilum* habitat suitability in Thailand

The model attained a total classification accuracy of 0.78. The results indicate that the Random Forest model yields dependable and resilient forecasts of habitat suitability. The integration of AUC and TSS values substantiates that the model possesses robust discriminatory capability under both threshold-independent and threshold-dependent assessment criteria. The optimal classification threshold identified using TSS maximization was 0.47, which represents the best trade-off between sensitivity and specificity rather than an arbitrary cutoff value.

Table 7: Performance evaluation metrics of the Random Forest model

Metric	Value
Accuracy	0.78
Sensitivity	0.80
Specificity	0.76
TSS	0.56
AUC	0.94
Brier score	0.05

3.1.1 Exploratory data analysis

The descriptive statistics of all predictor variables are presented in Table 7. The variables displayed diverse ranges and distributions, indicating the environmental variety throughout Thailand. Elevation and temperature-related variables had reasonably continuous distributions, whereas soil variables like $CaCO_3$ and sand content demonstrated substantial variability. Q–Q plots were created to evaluate the distributional properties of the predictors (Figure 5). Most variables had distributions that were nearly normal, while minor departures from normality were noted in NDVI and soil-related variables. Nonetheless, these aberrations were deemed non-problematic, as Random Forest is a non-parametric method that does not necessitate stringent normality assumptions. The exploratory study affirmed that the predictor variables were appropriate for inclusion in the Random Forest model without necessitating considerable data processing.

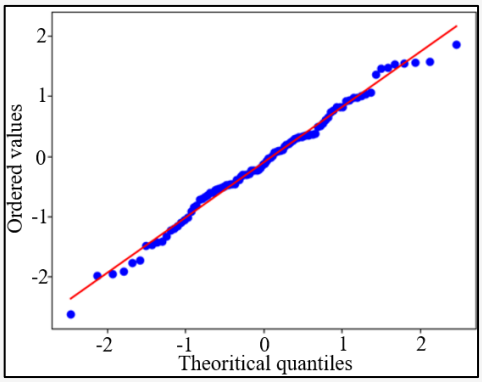

Figure 5: Q–Q plots of selected predictor variables used in the Random Forest model

3.1.2 Model performance evaluation

The efficacy of the Random Forest model was assessed utilizing both threshold-independent and threshold-dependent criteria. The model attained an AUC value of 0.94 (Figure 4), signifying exceptional discriminatory capability. An ideal classification threshold was established utilizing the maximum True Skill Statistic (TSS) to convert continuous probability outputs into binary classes. The selected threshold was 0.47, which provides a balance

between omission and commission errors. At this threshold, the confusion matrix (Table 6) indicated that the model attained a sensitivity of 0.80 and a specificity of 0.76, yielding a TSS value of 0.56. The model attained a total classification accuracy of 0.78. The results demonstrate that the model effectively differentiates between suitable and unsuitable settings for *Paphiopedilum* spp. The moderate TSS value signifies a judicious equilibrium between sensitivity and specificity, demonstrating the model's capacity to reduce both omission and commission errors.

3.2 Group-level Variable Importance of Predictor Variables

The relative contributions of predictor groups revealed that climatic variables accounted for 38.30% of the total model importance, followed closely by non-climatic environmental variables (37.40%), while spatial predictors contributed 24.30% of the explained importance (Table 8). These results indicate that the distribution of slipper orchids in Thailand is not determined solely by climatic conditions but rather by the combined effects of climate, terrain characteristics, vegetation structure, and geographic gradients. Similar patterns have been reported in other studies of montane plants and orchids, where local topography and microclimatic conditions strongly influence habitat suitability alongside regional climate gradients [1] and [20]. The substantial contribution of non-climatic variables highlights the ecological importance of terrain complexity and vegetation structure in shaping suitable habitats for *Paphiopedilum*. Mountainous regions often provide favorable microclimates characterized by cooler temperatures, higher humidity, and reduced thermal variability, conditions that are commonly associated with orchid habitats [5] and [6]. The combined contribution of non-climatic and spatial predictors (61.70%) also emphasizes the importance of incorporating landscape structure and geographic context when modelling species distributions at national spatial scales. Previous ecological modelling studies have similarly shown that including topographic and spatial variables significantly improves predictive performance beyond climate-only models [23]. Details are shown in Table 8.

Table 8: The overall contribution of each variable group

Variable group	Relative importance	Percentage (%)
Climatic	0.046	38.30
Non-Climatic	0.045	37.40
Spatial	0.029	24.30

3.3 Importance of Individual Environmental Predictors

The Random Forest model found environmental and spatial predictors as significant factors affecting the distribution of *Paphiopedilum* species. Elevation, temperature factors, and vegetation structure emerged as primary ecological determinants influencing species distribution patterns.

Subsequently, temperature-specific variables were examined: Temperature-related variables, such as the maximum temperature of the warmest month (TMAX) and the annual mean temperature (TEMP), were recognized as significant factors. Normalized Difference Vegetation Index (NDVI) vegetation structure was strongly influential as well, accounting for 9.68% of total model importance. Moisture-related parameters comprised precipitation in the driest quarter (PDRY), which accounted for 9.65% (mean decrease accuracy), and annual precipitation (PREC), which contributed 5.77%. Topographic variables also played a role in model predictions. Slope, which represents terrain structure, was the second most significant predictor, with a model significance of 5.44%. Soil-related variables had relatively low contributions. The calcium carbonate ($CaCO_3$) itself explained 4.82%, while sand fraction and forest type contributed 2.80% and 2.15%, respectively. Elevation proved to be a significant predictor, possibly indicating the species' affinity for particular altitudinal regions linked to lower temperatures, increased humidity, and diminished human disturbance. These conditions are recognized to facilitate the microhabitats essential for *Paphiopedilum* species. Temperature variables, especially yearly mean temperature and maximum temperature, are essential in ascertaining orchid distribution, as *Paphiopedilum* species exhibit heightened sensitivity to thermal conditions that affect physiological processes and reproductive success. NDVI underscores the significance of vegetation structure and canopy cover, which affect light availability, moisture retention, and microclimatic stability, essential elements for orchid survival.

The diminished significance of soil variables noted in this analysis may be affected by a scale discrepancy between predictor datasets. Soil properties, being intrinsically fine-scale and extremely diverse, were analyzed using raster layers aligned to a coarser spatial resolution (~1 km), in accordance with climatic variables. This discrepancy may conceal localized soil impacts crucial for the establishment of *Paphiopedilum*. Consequently, the little impact of soil factors should not be construed as a deficiency in ecological significance, but rather as

a constraint associated with data resolution and integration.

The strong influence of elevation and temperature related variables in the findings of this investigation are in accordance with previous ecological research on *Paphiopedilum* spp. and other montane orchids. Numerous slipper orchids are known to be found mostly in mountainous areas, where complicated topography, steady humidity, and lower temperatures provide ideal microhabitats for development and reproduction [5] and [6]. Elevation gradients frequently control local microclimates, which affect moisture availability, cloud cover, and temperature regimes, all of which are essential for orchid survival. In a similar vein, the NDVI's representation of vegetation structure is crucial for preserving stable microclimatic conditions and shaded forest understories that sustain orchid populations. NDVI readings may fluctuate seasonally due to alterations in vegetation phenology, thereby affecting their significance in the model. This study utilized NDVI data from the post-monsoon period, reflecting optimal vegetation conditions and serving as a reliable indicator of canopy structure. The application of NDVI from a singular time frame may inadequately represent seasonal fluctuations in vegetation dynamics. Future research should include multi-temporal NDVI datasets or seasonal composites to enhance the representation of temporal variability and refine the ecological interpretation of vegetation-related predictions. These results underscore the ecological sensitivity of *Paphiopedilum* orchids to environmental changes impacting forest structure and climate, as well as the significance of montane forest ecosystems as essential homes for these species.

While latitude (LAT) and longitude (LON) were recognized as significant predictors in the Random Forest model, however, These factors do not serve as direct ecological drivers; instead, they encapsulate extensive spatial structure and geographic gradients throughout the study area. Their relatively high importance may therefore reflect spatial autocorrelation, where nearby locations share similar environmental conditions and species occurrence patterns.

This study incorporated latitude and longitude to address extensive spatial processes, including dispersal limitations and the geographic aggregation of suitable habitats, which may not be entirely elucidated by environmental predictors alone. The persistent significance of environmental factors, such as elevation, temperature, and NDVI, suggests that essential ecological drivers remain strong even with the addition of spatial predictors. Future research must explicitly assess model sensitivity by comparing model performance and variable significance with and without spatial predictors, in addition to employing spatial cross-validation methods to more effectively differentiate spatial and environmental influences on species distribution patterns. The relative importance of predictor variables derived from the Random Forest model is presented in Table 9. Permutation importance values indicate the relative contribution of each predictor to model performance and should be understood in a comparative manner rather than as absolute indicators of ecological impact. Latitude (LAT) and elevation (DEM) were determined to be the most significant predictors, succeeded by longitude (LON) and temperature-related variables (TMAX and TEMP). While multiple predictors demonstrated analogous significance values, this indicates similar contributions to model performance rather than equal ecological functions. The minor variations noted among the highest-ranked variables indicate that numerous environmental factors together affect species distribution, rather than a singular predominant driver. This trend underscores the intricacy of ecological regulation; wherein numerous factors jointly influence habitat appropriateness.

Table 9: Individual variable contributions derived from the Random Forest model

No.	Variable	Importance	Group
1	LAT	0.015	Spatial
2	DEM	0.015	NonClimatic
3	LON	0.014	Spatial
4	TMAX	0.014	Climatic
5	TEMP	0.014	Climatic
6	NDVI	0.012	NonClimatic
7	PDRY	0.012	Climatic
8	PREC	0.007	Climatic
9	Slope	0.006	NonClimatic
10	$CaCO_3$	0.005	NonClimatic
11	Sand	0.003	NonClimatic
12	Forest Type	0.002	NonClimatic

3.4 Model Stability and Uncertainty

To evaluate the robustness of the RF model, the importance values were calculated from 20 separate runs of the same model, and both mean importance and standard deviation were determined for each predictor. Differences between single-run and mean importance values are expected due to the stochastic nature of the Random Forest algorithm. However, the overall ranking of key predictors remained largely consistent across runs, indicating the stability and robustness of the model results. This consistency suggests that the model is not overly sensitive to random variation in training data. The findings are shown in Table 10. The analysis indicated that multiple predictors showed consistently high mean

importance values in the cross-model-level iterations but with low standard deviation. In contrast, LON, DEM, TEMP, LAT and TMAX all featured consistently high importance values across simulation iterations. Low variability across these predictors suggests that the Random Forest model generated stable estimates of variable importance. This suggests that the main predictors affecting slipper orchid distribution remain stable over multiple models runs. Factors including precipitation of the driest quarter (PDRY) and slope showed moderate stability across model iterations, which may highlight their variable influence across the diverse environmental gradients of Thailand.

Table 10: Mean and standard deviation (SD) of Random Forest variable importance across 20 model runs

No.	Variable	Mean	SD
1	LON	0.014	0.002
2	DEM	0.014	0.001
3	TEMP	0.013	0.001
4	LAT	0.013	0.002
5	TMAX	0.013	0.002
6	NDVI	0.012	0.001
7	PDRY	0.011	0.002
8	Slope	0.007	0.002
9	PREC	0.006	0.001
10	$CaCO_3$	0.005	0.001
11	Forest type	0.003	0.001
12	Sand	0.003	0.001

In addition to the repeated model evaluation, the analysis of mean and standard deviation (SD) of variable importance further demonstrates the stability of the Random Forest model. As shown in Table 10, several predictors including longitude (LON), elevation (DEM), annual mean temperature (TEMP), latitude (LAT), and maximum temperature of the warmest month (TMAX) exhibit relatively high mean importance values with low standard deviations across the 20 model iterations. This indicates that these variables consistently contribute to model predictions regardless of resampling variation (Figure 6).

The narrow spread of SD values suggests that the Random Forest algorithm repeatedly identifies these predictors as key determinants of habitat suitability for *Paphiopedilum* species. Such stability indicates that the model is robust and that the identified predictors represent reliable environmental gradients influencing species distribution, which is a common advantage of Random Forest models in ecological applications [13][32] and [34]. In contrast, factors including the precipitation of the driest quarter (PDRY) and slope show moderate variability across model runs, suggesting that their influence may vary across environmental contexts within the study area. Meanwhile, edaphic variables including calcium carbonate ($CaCO_3$), sand fraction, and forest type exhibit both lower mean importance and slightly higher variability, indicating that their effects may operate primarily at finer spatial scales rather than at the national modelling extent. Overall, the mean ± SD analysis confirms that topographic gradients, thermal conditions, and vegetation structure represent the most stable environmental determinants of slipper orchid distribution across Thailand (Figure 6). To further evaluate whether differences in permutation importance among predictors were statistically significant, a one-way ANOVA was conducted. The results indicated no statistically significant differences among the top-ranked variables ($p > 0.05$), suggesting that these predictors contribute comparably to model performance.

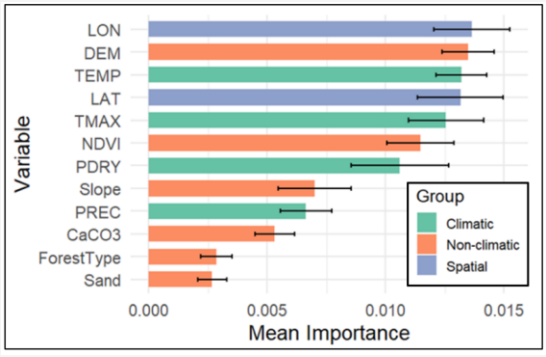

Figure 6: Variable importance (Mean ± SD) estimated by the Random Forest algorithm

3.5 Predicted Habitat Suitability

The persistent habitat suitability map produced by the RF model depict spatial probability of favourable environmental conditions for *Paphiopedilum* species throughout Thailand (Figure 7). Habitat appropriateness scores span from 0 to 1, with elevated values signifying more advantageous environmental conditions for species presence. The forecasted distribution pattern suggests that significant areas of Thailand have poor to moderate habitat appropriateness, especially in the central and northeastern regions. These regions are predominantly defined by lowland topography, intense agricultural practices, and comparatively elevated temperature conditions, which may restrict the presence of appropriate microhabitats for slipper orchids.

Conversely, places with high habitat appropriateness are predominantly located in the hilly areas of northern and western Thailand. The intricate topography, elevated altitudes, cooler climates, and generally undisturbed forest

ecosystems of these regions collectively create conducive environmental conditions for *Paphiopedilum* species. Mountain forests frequently offer stable microclimatic conditions, characterized by elevated humidity and diminished temperature fluctuations, which are crucial for the growth and survival of numerous orchid species [36].

Furthermore, numerous potentially appropriate habitats were identified in regions of southern Thailand, characterized by wet tropical forest ecosystems. Nevertheless, these appropriate ecosystems seem spatially fragmented and exist as dispersed patches rather than as continuous landscapes. Ecological fragmentation may restrict species migration and diminish genetic linkage among populations, potentially heightening the susceptibility of slipper orchid species to environmental changes and ecological disturbances. The anticipated habitat suitability pattern indicates that favorable conditions for *Paphiopedilum* species in Thailand are predominantly confined to mountainous forest ecosystems, indicating the importance of these habitats for orchid conservation and management.

3.6 Distribution of Habitat Suitability Classes

Habitat suitability values predicted by the Random Forest model were divided into three groups (low, moderate, and high appropriateness) based on anticipated suitability values, as shown in Figure 7.

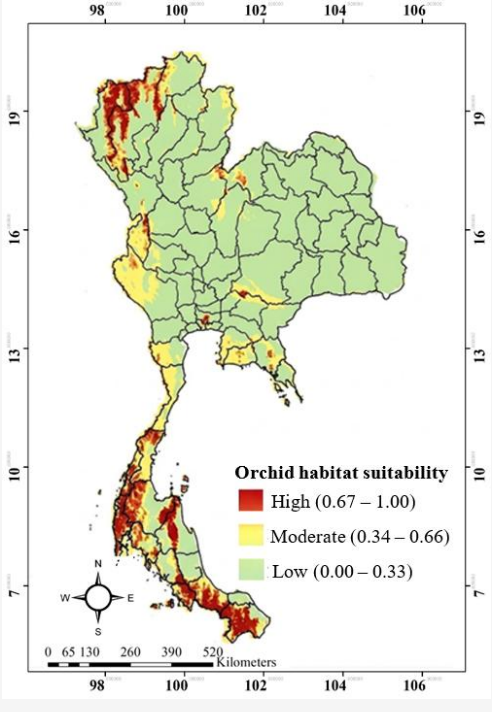

Figure 7: Habitat suitability map for *Paphiopedilum* spp. in Thailand

An equal interval system (0.00–0.33, 0.34–0.66, 0.67–1.00) was employed for visualization to illustrate relative variations in habitat appropriateness. This system is meant for mapping and should be taken as a simplified representation rather than precise ecological thresholds. The thresholds employed for mapping do not affect model evaluation, which was performed using a threshold optimized for True Skill Statistic (TSS). Future research should investigate alternate methodologies, such as natural breaks (Jenks) or threshold-based classification, to more effectively delineate ecological gradients. It is important to emphasize that the classification into low, moderate, and high suitability was used solely for visualization purposes. All model evaluation and interpretation were based on the data-driven optimal threshold derived from TSS maximization. This classification is intended solely for visualization and does not influence model evaluation or ecological interpretation. Moderately suitable habitats covered the largest proportion of the study area, accounting for 71.76% of the total area. Low suitability areas represented 25.04%, while high suitability habitats accounted for only 3.09% of the study region as shown in Figure 8. The relatively small proportion of highly suitable habitats indicates that environmental conditions supporting *Paphiopedilum* species are spatially limited.

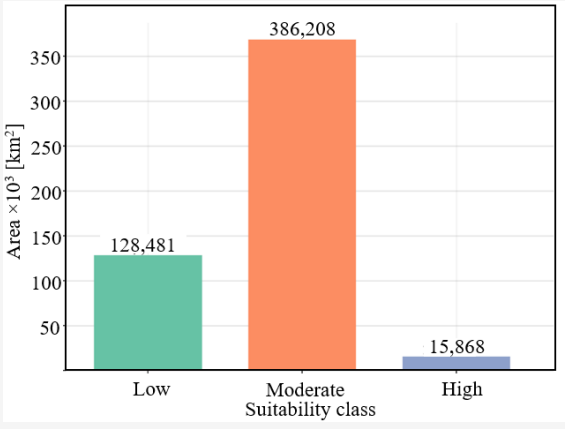

Figure 8: Area of habitat suitability classes derived from the Random Forest model

These habitats are primarily concentrated in mountainous forest landscapes where suitable microclimatic and ecological conditions occur. The spatial concentration of high-suitability habitats highlights the potential vulnerability of slipper orchid populations, as suitable environments may be restricted to specific ecological niches within the broader landscape.

4. Conclusion

This study assessed the significance of climatic, non-climatic, and geographical factors affecting the distribution of *Paphiopedilum* spp. in Thailand through a Random Forest modelling framework. The investigation, through the integration of multi-source geospatial datasets, found critical environmental variables linked to habitat suitability on a national scale. The findings demonstrate that the distribution of slipper orchids is influenced by the interplay of topographic gradients, climatic factors, vegetation structure, and spatial environmental patterns. Elevation, temperature-related variables, NDVI, and geographic coordinates were identified as the primary predictors, underscoring the significance of montane settings, thermal control, and forest structure in determining suitable habitats for *Paphiopedilum* species. The anticipated habitat suitability map indicated that highly appropriate habitats are primarily located in the mountainous areas of northern, western, and certain southern regions of Thailand, whereas extensive portions of the central and northeastern regions displayed reduced suitability.

Repeated model simulations revealed that the primary predictors were consistently steady, signifying resilient model performance. Nonetheless, certain restrictions must be recognized. The modelling methodology relied on a limited quantity of validated occurrence records, a coarse spatial resolution of environmental data, and randomly generated pseudo-absence points, which may inadequately represent the fine-scale ecological needs of slipper orchids. Moreover, spatial predictors may partially represent spatial autocorrelation instead of direct ecological processes, which could lead to misinterpretations of the ecological relationships affecting slipper orchids' distribution.

Notwithstanding these constraints, the study yields valuable insights into the extensive environmental contributing factors *Paphiopedilum* distribution in Thailand and establishes a scientific foundation for pinpointing priority regions for habitat protection and forthcoming field investigations. Future research must include spatial cross-validation, diverse pseudo-absence methodologies, higher-resolution environmental layers, and supplementary validation metrics to enhance ecological interpretation and predictive accuracy. A weakness of this work is to the geographical resolution of environmental variables (~1 km²), which may be too coarse for accurately predicting the distribution of *Paphiopedilum* species. Orchids exhibit a pronounced response to microhabitat conditions, encompassing subtle fluctuations in humidity, canopy cover, and substrate properties that may not be adequately represented at this level. Consequently, the model may insufficiently depict localized ecological circumstances, especially in varied mountainous terrains characterized by significant microclimatic change. The disparity between species ecology and the resolution of environmental data presents a prevalent challenge in species distribution modelling, potentially resulting in uncertainty in precise predictions.

Nonetheless, employing a 1 km resolution is suitable for national-scale study, since it facilitates the consistent integration of various environmental variables and encompasses extensive environmental gradients affecting species distribution. Subsequent research should integrate high-resolution environmental data and microhabitat characteristics to enhance ecological realism and forecast precision. While beyond our current scope, future research should apply spatial cross-validation to assess how latitude and longitude influence the importance of environmental predictors, ensuring a clearer distinction between spatial and environmental effects.

Acknowledgement

The authors would like to express their sincere gratitude to Mahidol University for academic support and research facilities. Special thanks are extended to the advisors and committee members for their valuable guidance and constructive comments throughout the study. The authors also acknowledge all organizations and institutions that provided environmental and species occurrence data used in this research.

References

[1] Arditti, J. and Ernst, R., (1992). *Fundamentals of Orchid Biology*. John Wiley & Sons Inc., New York.

[2] Chase, M. W., Cameron, K. M., Freudenstein, J. V., Pridgeon, A. M., Salazar, G., Van den Berg, C. and Schuiteman, A., (2015). An Updated Classification of Orchidaceae. *Botanical Journal of the Linnean Society*, Vol. 177; 151–174. https://doi.org/10.1111/boj.12234.

[3] Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. and Kent, J., (2000). Biodiversity Hotspots for Conservation Priorities. *Nature*, Vol. 403; 853–858. https://doi.org/10.1038/35002501.

[4] Cribb, P., (1998). *The Genus Paphiopedilum*. Natural History Publications, Kota Kinabalu, Malaysia.

[5] Bechtel, H., Cribb, P. and Launert, E., (1981). *The Manual of Cultivated Orchid Species*. MIT Press, Cambridge, MA, USA.

[6] Ticktin, T., (2023). Wild orchids: A Framework for Identifying and Improving Sustainable Harvest and Conservation Strategies. Biological Conservation, Vol. 282. https://doi.org/10.1016/j.biocon.2022.109816.

[7] Elith, J. and Leathwick, J. R., (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics, Vol. 40; 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

[8] Fay, M. F., (2018). Orchid Conservation: How Can We Meet the Challenges in the Twenty-First Century? *Botanical Studies*, Vol. 59. https://doi.org/10.1186/s40529-018-0232-z.

[9] Zurell, D., Franklin, J., König, C., Bouchet, P., Dormann, C., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J., Leitão, P., Park, D., Peterson, A., Rapacciuolo, G., Schmatz, D., Schröder, B., Serra-Diaz, J., Thuiller, W., Yates, K. L., Zimmermann, N. E. and Schröder, B., (2020). A Standard Protocol for Reporting Species Distribution Models. *Ecography*, Vol. 43; 1261–1277. https://doi.org/10.1111/ecog.04960.

[10] Morales-Linares, J., (2022). Habitat Diversity Promotes and Structures Orchid Communities in Tropical Ecosystems. *Flora*, Vol. 292. https://doi.org/10.1016/j.flora.2022.152180

[11] Guisan, A. and Thuiller, W., (2005). Predicting Species Distribution: Offering More than Simple Habitat Models. *Ecology Letters*, Vol. 8; 993–1009. https://doi.org/10.1111%2Fj.1461-0248.2005.00792.x.

[12] Olden, J. D., Lawler, J. J. and Poff, N. L., (2008). Machine Learning Methods Without Tears: A Primer for Ecologists. *Quarterly Review of Biology*, Vol. 83; 171–193. https://doi.org/10.1086/587826.

[13] Breiman, L., (2001). Random Forests. *Machine Learning*, Vol. 45; 5–32. https://doi.org/10.1023/A:1010933404324.

[14] Cutler, D. R., Edwards, T. C. and Beard, K. H., (2007). Random Forests for Classification in Ecology. *Ecology*, Vol. 88; 2783–2792. https://doi.org/10.1890/07-0539.1.

[15] Maxwell, A. E., Warner, T. A. and Fang, F., (2021). Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review. *International Journal of Remote Sensing*, Vol. 42; 2784–2817. https://doi.org/10.1080/01431161.2018.1433343.

[16] Peterson, A. T., Soberón, J. and Pearson, R. G., (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press, Princeton, NJ, USA.

[17] Royal Forest Department, (2021). Forest Resources Assessment of Thailand. *Royal Forest Department*, Bangkok, Thailand.

[18] Department of National Parks, Wildlife and Plant Conservation, (2022). Thailand Biodiversity Country Report. *Ministry of Natural Resources and Environment*, Bangkok, Thailand

[19] Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A., (2005). Very High-Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*, Vol. 25; 1965–1978. https://doi.org/10.1002/joc.1276.

[20] GBIF, (2024). GBIF Occurrence Download. [Online]. Available: https://www.gbif.org/ [Accessed: Jan. 5, 2026].

[21] Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S. E., Wisz, M. S. and Zimmermann, N. E., (2006). Novel Methods Improve Prediction of Species' Distributions from Occurrence Data. *Ecography,* Vol. 29; 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

[22] Pearson, R. G., Raxworthy, C. J., Nakamura, M. and Peterson, A. T., (2007). Predicting Species Distributions from Small Numbers of Occurrence Records. *Journal of Biogeography*, Vol. 34; 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

[23] Franklin, J., (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press, Cambridge, UK.

[24] Fick, S. E. and Hijmans, R. J., (2017). WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas. *International Journal of Climatology*, Vol. 37; 4302–4315. https://doi.org/10.1002/joc.5086.

[25] Dormann, C. F., Elith, J. and Bacher, S., (2013). Collinearity: A Review of Methods to Deal with it. *Ecography*, Vol. 36; 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

[26] Zuur, A. F., Ieno, E. N. and Elphick, C. S., (2010). A Protocol for Data Exploration to Avoid Common Statistical Problems. *Methods in Ecology and Evolution*, Vol. 1; 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.

[27] Belgiu, M. and Drăguţ, L., (2016). Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 114; 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[28] Barbet-Massin, M., Jiguet, F., Albert, C. H. and Thuiller, W., (2012). Selecting Pseudo-Absences for Species Distribution Models. *Methods in Ecology and Evolution*, Vol. 3; 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

[29] Fielding, A. H. and Bell, J. F., (1997). A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation*, Vol. 24; 38–49. https://doi.org/10.1017/S0376892997000088.

[30] Allouche, O., Tsoar, A. and Kadmon, R., (2006). Assessing the Accuracy of Species Distribution Models. *Journal of Applied Ecology*, Vol. 43; 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

[31] Pebesma, E., (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, Vol. 10; 439–446. https://doi.org/10.32614/RJ-2018-009.

[32] Araújo, M. B. and Guisan, A., (2020). Five (Or So) Challenges for Species Distribution Modelling. *Journal of Biogeography*, Vol. 47; 1419–1428. https://doi.org/10.1111/j.1365-2699.2006.01584.x.

[33] R Core Team, (2023). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria.

[34] Phillips, S. J., Anderson, R. P. and Schapire, R. E., (2006). Maximum Entropy Modeling of Species Geographic Distributions. *Ecological Modelling*, Vol. 190; 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

[35] Huang, S., Tang, M., Hupy, Y., Wang, Y. and Shao, G., (2021). A Commentary Review on The Use of NDVI in Ecological Studies. *Remote Sensing*, Vol. 13. https://doi.org/10.1007/s11676-020-01155-1.

[36] Fay, M. F. and Chase, M. W., (2021). Orchid Conservation: Current Status and Future Challenges. *Plants*, Vol. 10; 1–15. https://doi.org/10.1007/s10531-025-03196-6.

[37] Safdar, S., Bilal, M. and Khan, S., (2024). A Comprehensive Review of Spatial Distribution Modelling Techniques and Applications in Ecological Research. *Ecological Informatics*, Vol. 79. https://doi.org/10.1016/j.kjs.2024.100337.

[38] Karp, M. A. and Thorson, J. T., (2025). Applications of Species Distribution Modelling in Ecosystem Management and Conservation Planning. *ICES Journal of Marine Science*, Vol. 82. https://doi.org/10.1093/icesjms/fsaf024.