Incorrect and incomplete distribution data can mislead species modeling: a case study of the endangered Litsea auriculata (Lauraceae)

Chao Tan
Nanjing Forestry University
David Kay Ferguson
University of Vienna
Yong Yang
yangyong@njfu.edu.cn

Nanjing Forestry University

Research Article

Keywords: Conservation, Lauraceae, Litsea auriculata, MaxEnt, Species distribution modeling, Specimen identification

Posted Date: February 27th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3978669/v1

Additional Declarations: No competing interests reported.

Abstract

Global warming has caused many species to become endangered or even extinct. Describing and predicting how species will respond to global warming is one of the hot topics in the field of biodiversity research. Species distribution modeling predicts the potential distribution of species based on species occurrence records. However, it remains ambiguous how the accuracy of the distribution data impacts on the prediction results. To address this question, we used the endangered plant species *Litsea auriculata* (Lauraceae) as a case study. By collecting and assembling six different datasets of *Litsea auriculata*, we used MaxEnt model to perform species distribution modeling and then conducted comparative analyses. The results show that the distribution of *Litsea auriculata* is mainly in the Dabie Mountain region, southwestern Hubei and northern Zhejiang, and that mean diurnal temperature range (bio2) and temperature annual range (bio7) play important roles in the distribution of *Litsea auriculata.* Compared with the correct data, the dataset including misidentified specimens leads to a larger and expanded range in the predicted distribution area, whereas the species modeling based on the correct but incomplete data predicts a smaller and contracted range. According to the analysis of the local protection status of *Litsea auriculata*, we found that only about 23.38% of this species is located within nature reserves, so there is a large conservation gap. Our study suggests that the accurate distribution data is important for species modeling, and incomplete and incorrect data normally gives rise to misleading prediction results. In addition, our study also revealed the distribution characteristics and conservation gaps of *Litsea auriculata*, laying the foundation for the development of rational conservation strategies for this species.

1 Introduction

With global warming, populations of many species have been lost and fragmented, leading to an endangered status and even extinction of species (Power et al. 2019; Chase et al. 2020; Richards et al. 2020). A large number of species have become adapted to new distribution areas by modifying the pre-existing community composition and hence the ecosystem function (Babcock et al. 2019; Román-Palacios et al. 2020; Nielsen et al. 2021). Understanding the impact of future climate change on species potential habitats is important for the development of species conservation strategies (Austin et al. 2011; Hole et al. 2011; Moitz & Agudo 2013).

Species Distribution Models (SDMs) attempt to associate the distribution information of species with the corresponding environmental variables, establish models and predict the potential distribution of species in a certain area under specific spatial and temporal conditions in the future, which can quantify the regional and local distribution of species abundance at different scales (Guisan and Thuiller 2005; Araujo and Peterson 2012). SDMs mainly include a Generalized Linear Model (GLM), Classification and Regression Tree (CART), Random Forest (RF), Maximum Entropy (MaxEnt) etc. (Guo et al. 2020). The MaxEnt is one of the SDMs, based on extant species occurrence records and environmental data. It has the advantage of great accuracy, small sample size requirement, and good stability, and thus has become a widely used modeling approach (Wisz et al. 2008; Fitzpatrick et al. 2013; Merow et al. 2013; Morales et al. 2017; Wu et al. 2022). In recent years, there has been a steady increase in the literature on SDMs using the MaxEnt as a keyword in the Web of Science (Fig. 1). The use of SDMs makes it possible to predict the potential distribution of species in new space or time (Liu et al. 2022), and thus provides an important reference for species conservation.

SDMs are based on distribution site data and environmental factor data, so uncertainty in the location of species sampling sites will inevitably increase the uncertainty of modeling results (Guo et al. 2020). Specimen data have become an important data source for SDM predictions (Meineke et al. 2018). More than 3,000 herbaria in the world have collected over 400 million plant specimens (Thiers 2020). With the rapid digitization of plant specimens worldwide, specimen data have widely been used for different purposes, e.g. taxonomy, biogeography, phenology, and SDMs (Jaca et al. 2018; Jukonienė et al. 2018; Cámara-Leret et al. 2020; Meineke et al. 2018). However, it is worth pointing out that the herbarium collections and digitized specimens contain samples of cultivated plants far from their natural range as well as mis-identified material. There is a lack of quantitative description and research on how these misidentified and non-native specimen data have affected the results of SDMs.

Litsea auriculata is a deciduous tree species of the family Lauraceae. This species is characterized by scale-like exfoliating bark, large and auriculate leaves, long petioles, black ovoid fruits, and a cup-shaped receptacle. It has important economic and medicinal value, its wood has been used for furniture, while the fruits and roots have been employed as a traditional Chinese medicine (TCM) (Yang and Huang 1982). *Litsea auriculata* is sporadically distributed in a few mountainous areas at 500 – 1500m in Zhejiang, Anhui, Henan etc., and was listed as vulnerably endangered because of habitat loss and fragmentation (Fu and Jin 1992; Qin et al. 2017). As a result, it is important to investigate the conservation status of *Litsea auriculata* and identify any conservation gaps.

Species distribution modeling (SDM) should be based on complete sampling of accurately identified specimens, which is essential for understanding the suitable distribution area of species and for formulating reasonable conservation strategies (Costa et al. 2015; Fei and Yu 2016). Geng et al. (2017) conducted a study on community genetics and ecological niche modeling of *Litsea auriculata*, and predicted ecological niche shifts under different climate changes based on data from three populations in Tianmu Mountain of Zhejiang, Dabie Mountain of Anhui and Henan, and found that the habitat showed a trend towards contraction and decline in east-central China. However, the sampling range of this study is obviously inadequate, especially for the marginal areas of its distribution range. Based on specimens and literature data, Yang et al. (2018) documented the distribution of the species in Chun'an and Tiantai Counties in Zhejiang, Huoshan and She Counties in Anhui, Yingshan County and Shennongjia forest area in Hubei; Geng et al. (2017) did not include these localities. In addition, the Chinese Virtual Herbarium (abbreviated as CVH), the largest digitized herbarium data source, contains misidentified and cultivated specimens. However, it remains unclear how the incomplete sampling, misidentified and cultivated specimen data impact on the distribution modeling of this species.

In this study, we collected and collated six different datasets of *Litsea auriculata* and predicted each dataset using the MaxEnt, and compared the differences of the species distribution modeling results based on these different datasets. By doing this, we plan to answer the following three questions: 1) what are the impacts of misidentified and cultivated specimen data on the results of SDMs? 2) what are the differences between SDM results of inadequate sampling and complete and accurate datasets? 3) identify the conservation gap of the species based on our new species modeling results and indicate what action needs to be taken to conserve the species?

2 Materials and Methods

2.1 Data collection and processing

2.1.1 Distribution data of *Litsea auriculata*

The distribution data of *Litsea auriculata* were obtained from the Chinese Virtual Herbarium (CVH, https://www.cvh.ac.cn/), National Specimen Information Infrastructure (NSII, http://www.nsii.org.cn/2017/home.php), authoritative regional floras, and published papers (Sun 2014; Geng et al. 2017). We annotated the data source of each distribution record to generate different datasets. The distribution records were cross-checked for spelling errors. All the specimen records were visually identified by the corresponding author (Yong Yang), and the misidentified and cultivated records were labeled. Then, the collected data were further processed and separated into six datasets (see Table S1): dataset 1 (correct) including all the correctly identified records from herbarium specimens and literature; dataset 2 (cultivated) containing correctly identified and cultivated specimens; dataset 3 (misidentified) encompassing correctly identified and misidentified specimens but excluding cultivated specimens; dataset 4 (specimen) including only correctly identified specimens; dataset 5 (population) was collected from the literature, contained field population investigations (correctly identified but incomplete); dataset 6 (including all different sources) included all the distribution records of population investigations and herbarium data (correctly identified, misidentified and cultivated). We used Google Maps (http://maps.google.cn/) to obtain the geographic coordinates of the distribution records. We removed duplicate specimens and redundant records within the different datasets, before MaxEnt analysis and imported the distribution data into ArcGIS 10.2 to eliminate duplicate points, i.e., only one of the distribution records within 10 km was retained (Zhou et al. 2021).

2.1.2 Environment variable data

Altogether 19 environmental variable data at 2.5' resolution were downloaded from WorldClim (https://www.worldclim.org/) (see Table S2), including current climatic data (1970–2000) and future climate predictions. The future climatic data were based on the climate model of the Beijing Climate Center Climate System Model Version 1.1 (BCC-CSM 1.1), which was constructed under RCP 2.6, RCP 4.5, and RCP 8.5 for 2050 (average value over the period 2041–2060) and 2070 (average value over the period 2061–2080) for the three representative concentration pathways (RCPs) (Luo et al. 2009).

The climate layers were extracted using the software ArcGIS 10.2, and the extracted layers were converted to the ASCII format. In order to avoid influencing the final assessment of the model of high correlations between environmental variables (Luo et al. 2017), we conducted Pearson correlation analyses of 19 climatic variables for each period using the *cor* function of R software, and the climatic factors with r<|0.85| that were more closely related to species distribution were retained (Yan et al. 2017; Zhu et al. 2019). Finally, we performed principal component analyses (PCA) on the variables under current climatic conditions to identify the key drivers influencing the distribution of *Litsea auriculata.*

2.2 Potential distribution prediction using MaxEnt

Firstly, we imported the six distribution datasets (.CSV format) and climatic data (.ASCII format) for each period into MaxEnt 3.4.1 software for species ecological niche simulation. Secondly, different procedures for simulating the potential distribution were performed for datasets with different sample sizes. For data sets with fewer than 25 coordinate points, the Jackknife method was used for simulation evaluation. For species modeling, one of the coordinates was removed and the model was built based on the remaining n-1 coordinates, so that n models could be built and the optimal model selected for the MaxEnt ecological niche simulation. For data sets with more than 25 available coordinate points, 75% of the species distribution data was set as the training set and 25% as the test set, the number of operational iterations was set to 10, and the rest was used as default values (Pearson et al. 2007; Zhou et al. 2021). The area under curves (AUC) with receiver operator characteristic (ROC) was used to evaluate the reliability of the simulation results (Guo et al. 2019). The range of AUC values was 0 to 1, the closer to 1 indicating the higher reliability of the simulation. The simulation result was considered to be very accurate when the AUC value was between 0.9 and 1, accurate when the AUC was 0.8 – 0.9, average when the AUC was between 0.7 and 0.8, and unreliable when the AUC result was less than 0.7 (Elith et al. 2016; Jiang et al. 2016). Finally, the simulation results of MaxEnt were entered into ArcGIS 10.2 software and transformed into raster layers for visualization, and the natural breaks method was selected to calculate the fitness index P. Based on previous studies, P>0.75 was used as a hotspot for species survival (Shi et al. 2022), and the proportion of the area in different distribution data types was calculated.

2.3 Calculating hotspots in protected areas

To describe and evaluate the local conservation status of *Litsea auriculata*, we assembled 2569 nature reserves (including 440 national nature reserves and 2,129 provincial and county nature reserves) established during 1956 – 2021 (Zhang et al. 2015). In ArcGIS 10.2, the base map data of China's nature reserves superimposed on the samples were used to calculate the area of the contemporary hotspot area located within the reserve, and to evaluate the protection efficiency of *Litsea auriculata*.

3 Results

3.1 Current distribution pattern of *Litsea auriculata*

Six distribution datasets were assembled in this study. Dataset 4 contains 16 records, and was based on herbarium specimen data from CVH and NSII. Dataset 5 was collected from the literature, and contained 9 records. Dataset 1 was an integration of dataset 4 and dataset 5, and consisted of a total of 18 records after removing duplicate records. Both dataset 2 and dataset 3 were assembled using specimen data from CVH and NSII, each containing 22 records. Dataset 6 was an integration of dataset 2, dataset 3, dataset 4, and dataset 5, and contained a total of 26 records after deleting duplicate records.

According to the correct and complete dataset (dataset 1), *Litsea auriculata* was distributed in Dabie Shan at the border of Henan and Anhui, Qingliang Mountain at the border of Anhui and Zhejiang, Daming Mountain in Zhejiang, and Nanzhao County of Henan and Shennongjia forestry district in Hubei (Fig. 2). This species was introduced to botanical gardens outside its native range for the purpose of ex situ conservation, e.g. Ming Xiaoling Mausoleum in Jiangsu, Hangzhou Botanical Garden in Zhejiang, Lushan Botanical Garden in Jiangxi, and Kunming Botanical Garden in Yunnan (Fig. 2). Wrong identification records expanded the distribution range of the species, e.g. Chongyi County in Jiangxi, Fengkai County in Guangdong, Jiangshan County in Zhejiang, and Sandu Shui Autonomous County of Guizhou (Fig. 2).

The six datasets were screened for environmental variables based on Pearson correlation analyses. The results show that dataset 1 and dataset 3 each retained six climate factors in the final MaxEnt model analyses. The remaining datasets retained five climate factors in the final model analysis, respectively (Table 1).

Table 1
Screened environmental variables of different datasets for the final MaxEnt model analysis. Details of climate variables see Table S2.

Type	bio1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
datase 1		*		*			*				*				*			*	
datase 2				*		*	*								*			*	
datase 3				*	*		*		*						*			*	
datase 4			*	*			*				*							*	
datase 5	*	*		*											*			*	
datase 6				*		*	*								*			*	

3.2 Spatial pattern and driving factors of potential distribution areas of various data

Based on the assembled distribution datasets and environmental data, the potential geographical distribution area of this species was simulated using the optimal MaxEnt model. The results show that the AUC value of the simulated curves of all six datasets was greater than 0.994, indicating that the prediction results of the model are very reliable (see Table S3).

The potential distribution patterns based upon different datasets were significantly different under current climatic condition. The suitable areas predicted for *Litsea auriculata* based on the correct dataset (dataset 1) were mainly distributed in Dabie Mountain, Huangshan Mountain and southwestern Hubei, and a small area in Zhejiang (Fig. 3a). Under-sampled datasets predicted distribution areas showing minor differences from the correct dataset (dataset 1). Compared with the predicted result of dataset 1, the extent of the fitness zone based upon the specimen dataset (dataset 4) extended in easterly and westerly directions and shrunk in the middle part (Fig. 3d), while the suitable area based upon the population dataset (dataset 5) shrunk gradually from the periphery to the middle (Fig. 3e). The difference between the suitable areas based upon the inaccurate dataset (datasets 2, 3 & 6) and the correct dataset (dataset 1) was rather obvious, and the potential distribution areas based upon these three datasets were widely distributed and extended in all directions, throughout the middle and lower reaches of the Yangtze River (Fig. 3b,c,f). Besides, under the 2050s and 2070s RCP 2.6/4.5/8.5 climate scenarios, the suitable areas based upon these different datasets were basically consistent with those under contemporary conditions (see Figure S1).

The predicted hotspot areas based upon different datasets showed distinct trends under various climatic conditions (Fig. 4). The correct but incomplete datasets (datasets 4 & 5) displayed minor differences from the correct dataset (dataset 1), ranging from 0.01–0.54%. The largest hotspot area anomaly was under the 2050s RCP 2.6 condition, where the population dataset (dataset 5) differs from the correct dataset (dataset 1) by 0.54% with an area of 51,900 km². The smallest hotspot area occurred under multiple climate scenarios, the suitable area based on the specimen dataset (dataset 4) differed from that based on the correct dataset (dataset 1) by 0.01% with only 1,000 km² under the 2050s RCP 4.5/8.5 climatic conditions. The same result appears in the 2070s RCP 8.5, with the specimen dataset (dataset 4) and population dataset (dataset 5) differing by 1,000 km² from the correct dataset (dataset 1).

The incorrect dataset (datasets 2, 3 & 6) and the correct dataset (dataset 1), on the other hand, exhibited a large difference of 0.03%–0.88%. The largest hot spot area discrepancy value occurred in the misidentified dataset (dataset 3) for the 2070s RCP 8.5 with 0.88% and an area of 82,600 km². The smallest area gap of 2,900 km² occurred in all datasets (dataset 6) under 2050s RCP 2.6. In addition, the maximum hotspot area difference in all climatic environments occurred in the predicted fitness zones of the misidentified dataset (dataset 3), except for 2070s RCP 2.6 which materialized in the cultivated dataset (dataset 2) (Table 2).

Table 2
The proportion of hotspot areas in different datasets (hotspots/selected regions).

Type	Present	2050s			2070s		
		RCP2.6	RCP4.5	RCP8.5	RCP2.6	RCP4.5	RCP8.5
dataset 1	0.16%	0.70%	0.14%	0.07%	0.39%	0.15%	0.08%
dataset 2	0.30%	0.29%	0.30%	0.31%	0.92%	0.25%	0.18%
dataset 3	0.42%	0.92%	0.50%	0.57%	0.91%	0.46%	0.96%
dataset 4	0.09%	0.48%	0.13%	0.08%	0.12%	0.12%	0.07%
dataset 5	0.11%	0.16%	0.07%	0.10%	0.11%	0.12%	0.07%
dataset 6	0.38%	0.67%	0.47%	0.30%	0.53%	0.43%	0.19%

3.3 PCA of *Litsea auriculata* different datasets under current climatic condition

The contribution of environmental variables varied when conducting PCA studies based on different datasets under current climatic condition (see Table S4). Mean diurnal temperature range (bio2) and temperature annual range (bio7) played a decisive role in the correct dataset (dataset 1) of *Litsea auriculata* (Fig. 5a). bio7 and Isothermality (bio3) had the largest impact on the specimen dataset (dataset 4) prediction (Fig. 5d), while bio2 and precipitation seasonality (bio15) determined the distribution of the population dataset (dataset 5) (Fig. 5e). In the incorrect datasets (datasets 2, 3 & 6), the two most important determinants for the distribution of cultivated (dataset 2) and all recorded datasets (dataset 6) were bio7 and temperature seasonality (bio4) (Fig. 5b,d), while the distribution of the misidentified dataset (dataset 3) was limited by mean temperature of the driest quarter (bio9) and bio7 (Fig. 5c).

3.4 Distribution and conversation status of *Litsea auriculata* using different datasets under current climatic condition

Under contemporary climatic conditions, the hotspots and protection status predicted based on the different datasets displayed great discrepancies. The hotspots based on the correct dataset (dataset 1) were mainly distributed in Dabie and Huangshan Mountains, with small stands in southwestern Hubei and Zhejiang. The total area was 15,400 km², of which 3,600 km² (23.38%) was located in nature reserves (Fig. 6a).

The range of hotspots predicted by the inaccurate dataset (datasets 2, 3 & 6) displayed a certain degree of expansion compared with the correct dataset (dataset 1). The hotspots of the cultivated dataset (dataset 2) were concentrated in the Dabie and Huangshan Mountains, with a small area in Hunan, a total area of 28,800 km², of which 2,600 km² (9.03%) was in a protected area (Fig. 6b). The range of hotspots predicted by the inclusive dataset (dataset 6) was similar to that of the cultivated dataset, with additional distribution areas in southwestern Zhejiang; the total area of the hotspot range was 36,500 km², only 3,300 km² (9.04%) was located in a protected area (Fig. 6c). The misidentified dataset (dataset 3) predicted the largest hotspot area of 40,400 km², which formed a dense area in southwestern Hubei and northwestern Hunan compared with the cultivated dataset, and extended outwards from the Dabie and Huangshan Mountains, with only 5,300 km² (13.18%) located in a protected area (Fig. 6f).

The distribution range of hotspot regions predicted by the correct but incomplete dataset (datasets 4 & 5) was similar to the correct dataset (dataset 1), and showed an overall contraction. The hotspot areas of the population dataset (dataset 4) contracted towards the central area of the correct dataset (dataset 1), possessed a total area of 11,500 km²with only 1,800 km²(15.65%) in nature reserves (Fig. 6e). Species modeling based on the specimen dataset (dataset 5) showed a shrinking trend in the hotspots and a scattered occurrence in southwestern Hubei, the total area covering ca. 8,600 km² with only 1,700 km² (19.77%) hotspot area in nature reserves (Fig. 6d).

Table 3
The hotspot areas, area and proportion of the hotspots in nature reserves according to species modeling using different datasets under contemporary climatic conditions. (Unit: km²)

Type	Hotspot areas	Predicted areas in nature reserves	Proportion
dataset 1	15,400	3,600	23.38%
dataset 2	28,800	2,600	9.03%
dataset 3	40,400	5,300	13.18%
dataset 4	11,500	1,800	15.65%
dataset 5	8,600	1,700	19.77%
dataset 6	36,500	3,300	9.04%

4 Discussion

4.1 Importance of accurate identification and complete species distribution records for species modeling

Distribution data is the basis for species modeling predictions. Kadmon et al. (2004) conducted a comparative study on the distribution modeling of 149 woody plant species in Israel, which revealed that data biases can reduce the accuracy of species modeling, the same conclusion was found by Kramer-Schadt et al. (2013). Raes & ter Steege (2007) performed a null model test on species modeling and found that modeling with incorrect distribution data showed significantly different results from the correct data set, demonstrating the impact of data bias on species modeling, which was further corroborated by Wolmarans et al. (2010) and Chen et al. (2015). In this study, we compared predictions based on distribution records containing cultivated/misidentified records (datasets 2, 3 & 6) with those based on correctly identified and complete natural distribution records (dataset 1). Our results indicate that the dataset containing misidentified specimens can result in expansion of the fitness areas, thus significantly reducing the accuracy of the model. We compared the prediction results based on the distribution dataset containing cultivated records with those of the correctly identified complete natural distribution records, and found that the suitable distribution area expands greatly from the center to the surrounding area. This indicates that the modeling accuracy decreases with increasingly biased data. Our comparative study of species modeling results based on incomplete natural distribution records (datasets 4 & 5) and correctly identified, complete natural distribution records (dataset 1) suggests that the suitable area showed a conspicuous contraction trend with a very narrow distribution. Species modeling predictions based on such misidentified and inaccurate specimen data can arrive at misleading conclusions.

With the rapid development of digital cameras, computers, and internet information technology, a large number of herbarium specimens throughout the world have been digitized and are available for biodiversity studies (Meineke et al. 2018; Davis 2023). By June 2023, 0.24 billion specimens had been included in the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/) and 11.59 millions of specimen data deposited in the Australian Biological Atlas / Atlas of Living Australia (ALA, https://www.ala.org.au/). The National Plant Specimen Resource Center (NPSRC, http://www.cvh.ac.cn/), the largest digital plant specimen integration platform in China, has collected 8.27 million digitized plant specimens. National Specimen Information Infrastructure (NSII) contains about 16.45 million digital plant specimens. These digitized specimens have become important sources for research in ecology, biogeography, phenology, and conservation biology (Merow et al. 2016; Nualart et al. 2017; Jones and Daehler 2018; Herbling 2022; Yang et al. 2022; Lee et al. 2022; Davis 2023). However, over 50% of the herbarium specimens were not correctly identified (Goodwin et al. 2015). Digitized specimens thus contain lots of identification errors and cultivated records, and are the main source of erroneous data in species modeling. Incorrect distribution information often leads to severe range deviations and obscures the true species model (Orr et al. 2021). As a result, it is necessary to remove and correct the misidentified records and cultivated records before conducting species model predictions.

Because published floras record older data and often contain incomplete information, the integration of floras cannot resolve the problem of data completeness. In this study, we found that the Flora of China records the distribution of *Litsea auriculata* in Tianmu Mountain and Tiantai Mountain in Zhejiang and She County in Anhui (Yang and Huang 1982), and misses many other distribution localities. Our new inventory in this study has added the records of *Litsea auriculata* in Hubei and Henan, and Chun'an County in Zhejiang. The Flora of Anhui is comprehensive at the county level, but remains ambiguous regarding the distribution below the county level. The Jiangxi Seed Plant List contains an incorrect record of *Litsea auriculata*, which originates from misidentified digitized specimens (Liu et al. 2010). The distribution information in these botanical catalogs is fragmentary and cannot be used directly for species modeling, and needs to be verified and integrated. Only when complete and accurate data are available we can obtain valuable research results, which can help understand the distribution characteristics of species and provide important references for biodiversity conservation.

Specimens comprise the primary source of species distribution data, and should be correctly identified by taxonomists before utilization. Correct identification is fundamental not only for species distribution modeling, but also for biodiversity conservation. However, taxonomy as a traditional discipline is handicapped in the assessment and evaluation system of many different research institutions (Ma 2014). Most research funding has been deployed in more fashionable and advanced research areas, e.g. genome sequencing, making it difficult to train traditional taxonomists. As a result, no taxonomists work in the herbaria to correct the misidentified specimens. To overcome this drawback, it is necessary to promote traditional taxonomy and maintain a permanent taxonomic research team.

4.2 Potential distribution and conservation assessment based on accurate identification and complete dataset of *Litsea auriculata*

In this study, we established a reliable potential distribution area for *Litsea auriculata* based on an accurately identified and complete dataset (dataset 1). The modeling results show that, compared with other plants of the Lauraceae family (Zheng et al. 2018), the distribution range of this species is generally northerly and is currently located mainly on montane forest slopes in the mid-latitudes of central-eastern China. The predicted distribution is similar to the distribution characteristics of gymnosperm species (Tang et al. 2006; Xie et al. 2021). With global warming in the future, the suitable distribution area of *Litsea auriculata* will tend to contract, and eventually decrease in the central-eastern part of China, which corroborates a previous study (Geng et al. 2017). The predicted hotspot areas using accurately identified and complete datasets (dataset 1) under the contemporary climate shifted southwards compared to Geng et al. (2017). This difference may be caused by the bias of distribution data, as Geng et al. (2017) did not fully record the distribution of the species in southern regions such as Anhui and Zhejiang.

The potential distribution trend of *Litsea auriculata* shows a clear mismatch with subtropical broadleaved evergreen forest plants. Previous studies have suggested that subtropical broadleaved evergreen forest species will expand northwards and eastwards under future climatic conditions (Hu et al. 2017; Lim et al. 2018; Wu et al. 2016). The potential distribution ranges of *Litsea auriculata* do not vary significantly across time, with an overall range of only 0.09%–0.54%, the only local expansion and contraction occurring in some mountains and plains at the edges of the subtropical broadleaved evergreen forests. Coincidentally, a similar pattern was also found in a study of the genus *Cinnamomum* (Zhou et al. 2021). In addition, as in many gymnosperms, *Litsea auriculata* may have survived by elevational shifts during the late Quaternary glacial oscillations (Cun and Wang 2015).

The survival of *Litsea auriculata* is at least partially attributable to its habitat dilemma. Previous studies have shown that the genetic structure of *Litsea auriculata* continues to diverge and expand, forming small-scale populations (Sun 2014). Increased random genetic variation, high levels of inbreeding and reduced gene numbers, combined with a progressively warmer climate, have led to a dramatic decline in the distribution area of this species (Geng et al. 2017). In our study, the predicted results based on an accurately identified and complete dataset (dataset 1) for hotspot areas of *Litsea auriculata* under contemporary climatic conditions show that the species continues to spread in all directions in the future, with increased fragmentation, a gradual reduction in living space, and a further decrease in area, which is consistent with the results of previous studies. Besides, the narrow and concentrated distribution area has increased the threat level of *Litsea auriculata* (Qin et al. 2017), and irreversible damage will occur if these small areas are disturbed.

Species distribution models can suggest the chances of survival of endangered plants and facilitate the development of targeted *in situ* conservation measures (Aguilar-Soto et al. 2015). In this paper, habitat prediction in combination with an analysis of Chinese nature reserves, indicates that only 23.38% of *Litsea auriculata* is currently located in nature reserves, so a large conservation gap remains. The areas outside the nature reserves are mainly located in southern Anhui, west-central and east-central Zhejiang. These areas have suffered from severe deforestation, habitat loss and habitat fragmentation (Wei and Jiang 2012), which may have lead to a significant decrease in the number and population size of *Litsea auriculata.* The area of the species within the nature reserve will gradually shrink under future warming scenarios, and may even deviate excessively from the reserve in the 2070s RCP 2.6 scenario (Table 4), thus greatly increasing the threat level. Therefore, in the face of such a situation, a protected area should be established for *Litsea auriculata*, and special staff should be assigned to protect the forest land, prohibit indiscriminate logging practices, and reduce human interference. According to previous studies, we found that a large number of threatened gymnosperms also survive in the distribution area of *Litsea auriculata* (Lü et al. 2018; Xie et al. 2021), so it is crucial to strengthen the protection of these areas for other threatened plants as well. In addition, because the genetic differentiation among populations of *Litsea auriculata* is large and gene flow is low (Sun 2014), it would be beneficial to increase the level of genetic diversity of *Litsea auriculata* if a sufficient number of individuals within all populations could be selected for intensive translocation and conservation.

Table 4
Predicted hotspot area of *Litsea auriculata* based on correct dataset (dataset 1) and area located within the protected area. (Unit: km²)

Period	Hotspot areas	Nature reserves areas	Proportion
Current	15,400	1,700	23.38%
2050s RCP2.6	67,300	6,400	9.06%
2050s RCP4.5	13,500	2,200	16.30%
2050s RCP8.5	6,700	1,300	19.40%
2070s RCP2.6	37,500	2,900	7.73%
2070s RCP4.5	14,400	2,300	15.97%
2070s RCP8.5	7,700	1,200	15.58%

5 Conclusion

It remains ambiguous how the identification errors, cultivated collections and data incompleteness impact on species distribution modeling. We assembled six datasets and made a comparative study here. We show that misidentification, cultivated specimen data, and data incompleteness all have significant impacts on species modeling prediction results. We identified new areas of potential distribution of *Litsea auriculata* based on correctly identified and more complete datasets, revealed that the current main distribution range of *Litsea auriculata* is located in the mountainous areas of the middle and lower reaches of the Yangtze River, with a tendency to contraction in future climate change scenarios. In addition, our assessment of the conservation status of *Litsea auriculata*, reveals that currently about 23.38% of the suitable areas for the species have been protected in nature reserves, so there are still relatively large conservation gaps. The resulting information can be used to support management, conservation, and recovery plans for *Litsea auriculata*.

Declarations

Funding

This work was supported by the National Natural Science Foundation of China [32270217 and 31970205] and the *Metasequoia* funding of the Nanjing Forestry University. The authors declare no competing interests.

Author Contribution

Chao Tan and Yong Yang wrote the main manuscript text and Chao Tan prepared figures 1-6. All authors reviewed the manuscript

Acknowledgements

We are grateful to the contributors to the Chinese Virtual Herbarium.

Data Availability Statement

All data used in the study are included in this paper are available in the supporting datasets.

References

1. Aguilar-Soto V, Melgoza-Castillo A, Villarreal-Guerrero F, Wehenkel C, Pinedo-Alvarez C (2015) Modeling the potential distribution of *Picea chihuahuana* Martínez, an endangered species at the sierra madre occidental. Mexico Forests 6:692–707. https://doi.org/10.3390/f6030692
2. Araujo MB, Peterson AT (2012) Uses and misuses of bioclimatic envelope modeling. Ecology 93:1527–1539. https://doi.org/10.1890/11-1930.1
3. Austin MP, Van Niel KP (2011) Improving species distribution models for climate change studies: variable selection and scale. J Biogeogr 38:1–8. https://doi.org/10.1111/j.1365-2699.2010.02416.x
4. Babcock RC, Bustamante RH, Fulton EA et al (2019) Severe continental-scale impacts of climate change are happening now: extreme climate events impact marine habitat forming communities along 45% of Australia's coast. Front Mar Sci 6:411. https://doi.org/10.3389/fmars.2019.00411
5. Camara-Leret R, Frodin DG, Adema F et al (2020) New Guinea has the world's richest island flora. Nature 584:579–583. https://doi.org/10.1038/s41586-020-2549-5
6. Chase JM, Blowes SA, Knight TM, Gerstner K, May F (2020) Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature 584:238–243. https://doi.org/10.1038/s41586-020-2531-2
7. Chen SB, Slik JWF, Mao LF, Zhang J, Sa RL, Zhou KX, Gao JX (2015) Spatial patterns and environmental correlates of bryophyte richness: sampling effort matters. Biodivers Conserv 2:593–607. https://doi.org/10.1007/s10531-014-0838-8
8. Costa H, Foody GM, Jimenez S, Silva L (2015) Impacts of species misidentification on species distribution modeling with presence-only data. ISPRS Int J Geo-Inf 4:2496–2518. https://doi.org/10.3390/ijgi4042496

9. Cun YZ, Wang XQ (2015) Phylogeography and evolution of three closely related species of *Tsuga* (hemlock) from subtropical eastern Asia: further insights into speciation of conifers. J Biogeogr 42:315–327. https://doi.org/10.1111/jbi.12421
10. Davis CC (2023) The herbarium of the future. Trends Ecol Evol 38:412–423. https://doi.org/10.1016/j.tree.2022.11.015
11. Elith J, Graham CH, Anderson RP et al (2016) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x
12. Fei S, Yu F (2016) Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. Landsc Ecol 31:31–42. https://doi.org/10.1007/s10980-015-0272-7
13. Fitzpatrick MC, Gotelli NJ, Ellison AM (2013) MaxEnt versus MaxLike: empirical comparisons with ant species distributions. Ecosphere 4:1–15. https://doi.org/10.1890/ES13-00066.1
14. Fu LG, Jin JM (1992) China Plant Red Data Book: Rare and Endangered Plants Volume 1. Science, Beijing
15. Geng QF, Sun L, Zhang PH, Wang ZS, Qiu YX, Liu H, Lian CL (2017) Understanding population structure and historical demography of *Litsea auriculata* (Lauraceae), an endangered species in east China. Sci Rep 7:1–16. https://doi.org/10.1038/s41598-017-16917-x
16. Goodwin ZA, Harris DJ, Filer D, Wood JRI, Scotland RW (2015) Widespread mistaken identity in tropical plant collections. Curr Biol 25. https://doi.org/10.1016/j.cub.2015.10.002. R1057-R1069
17. Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009. https://doi.org/10.1111/j.1461-0248.2007.01044.x
18. Guo YL, Li X, Zhao ZF, Nawaz Z (2019) Predicting the impacts of climate change, soils and vegetation types on the geographic distribution of *Polyporus umbellatus* in China. Sci Total Environ 648:1–11. https://doi.org/10.1016/j.scitotenv.2018.07.465
19. Guo YL, Zhao ZF, Qiao HJ et al (2020) Challenges and development trend of species distribution model. Adv Earth Sci 35:1292–1305. https://doi.org/10.11867/j.issn.1001-8166.2020.110
20. Herbling JM (2022) Herbaria as big data sources of plant traits. Int J Plant Sc 183:87–118. https://doi.org/10.1086/717623
21. Hole DG, Huntley B, Arinaitwe J et al (2011) Toward a management framework for networks of protected areas in the face of climate change. Conserv Biol 25:305–315. https://doi.org/10.1111/j.1523-1739.2010.01633.x
22. Hu WQ, Wen NL, Xiao Y, Zhong RZ, Ru ZZ (2017) Geographic distribution and potential distribution estimation of *Machilus breviflora*. Guangdong Agri Sci 44:82–85. https://doi.org/10.1111/j.1523-1739.2010.01633.x
23. Jaca TP, Boatwright JS, Moteetee AN (2018) Taxonomic studies of the genus *Rhynchosia* Lour. (Phaseoleae, Fabaceae) in South Africa: A review of section *Chrysoscias*. S Afr J Bot 117:119–133. https://doi.org/10.1016/j.sajb.2018.05.012
24. Jiang XL, Deng M, Li Y (2016) Evolutionary history of subtropical evergreen broad-leaved forest in Yunnan Plateau and adjacent areas: An insight from *Quercus schottkyana* (Fagaceae). Tree Genet Genomes 12:1–12. https://doi.org/10.1007/s11295-016-1063-2
25. Jones CA, Daehler CC (2018) Herbarium specimens can reveal impacts of climate change on plant phenology: a review of methods and applications. PeerJ 6:e4576. https://doi.org/10.7717/peerj.4576
26. Jukoniene I, Rasimavicius M, Rickiene A, Subkaite M (2018) S.B. Gorski's bryological collection in the herbarium of Vilnius University. Acta Soc Bot Pol 87:3588. https://doi.org/10.5586/asbp.3588

27. Kadmon R, Farber O, Danin A (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 14:401–413. https://doi.org/10.1890/02-5364
28. Kramer-Schadt S, Niedballa J, Pilgrim JD et al (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. Divers distrib 19:1366–1379. https://doi.org/10.1111/ddi.12096
29. Lee BR, Miller TK, Rosche C, Yang Y, Heberling M, Kuebbing SE, Primark RB (2022) Wildflower phenological escape differs by continent and spring temperature. Nat Commun 13:7157. https://doi.org/10.1038/s41467-022-34936-9
30. Lim CH, Yoo S, Choi Y, Jeon S, Son Y, Lee WK (2018) Assessing climate change impact on forest habitat suitability and diversity in the Korean Peninsula. Forests 9:259. https://doi.org/10.3390/f9050259
31. Liu CL, Wolter C, Courchamp F, Roura-Pascual N, Jeschke JM (2022) Biological invasions reveal how niche change affects the transferability of species distribution models. Ecology 103: e3719. https://doi.org/0.1002/ecy.3719
32. Liu ZL, Zhang ZX, Liao WM (2010) Seed Plant List of Jiangxi. China Forestry Publishing House, Beijing
33. Lü LS, Cai HY, Yang Y, Wang ZH, Zeng H (2018) Geographic patterns and environmental determinants of gymnosperm species diversity in China. Biodivers Sci 26(11):1133–1146. https://doi.org/10.17520/biods.2018098
34. Luo X, Hu QJ, Zhou PP, Zhang D, Wang Q, Abbott RJ, Liu JQ (2017) Chasing ghosts: Allopolyploid origin of *Oxyria sinensis* (Polygonaceae) from its only diploid congener and an unknown ancestor. Mol Ecol 26:3037–3049. https://doi.org/10.1111/mec.14097
35. Luo Y, Wu TW (2009) The development and progress of climate system model BCC-CSM. //International Conference on Earth Science and Technology Abstracts in 2009 by Chinese Meteorological Society. Chinese Meteorological Society, Beijing
36. Ma JS (2014) Current status and challenges of Chinese plant taxonomy. Chin Sci Bull 59:510–521. https://doi.org/10.1360/972013-320
37. Meineke EK, Davies TJ, Daru BH, Davis CC (2018) Biological collections for understanding biodiversity in the Anthropocene. Philos T R Soc B 374:20170386. https://doi.org/10.1098/rstb.2017.0386
38. Merow C, Allen JM, Aiello-Lammens M, Silander JA Jr (2016) Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. Global Ecol Biogeogr 25:1022–1036. https://doi.org/10.1111/geb.12453
39. Merow C, Smith MJ, Silander JA Jr (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36:1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x
40. Morales NS, Fernández IC, Baca-González V (2017) MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. PeerJ 5:e3093. https://doi.org/10.7717/peerj.3093
41. Moritz C, Agudo R (2013) The future of species under climate change: resilience or decline? Science 341:504–508. https://doi.org/10.1126/science.1237190
42. Nielsen ES, Henriques R, Beger M, von der Heyden S (2021) Distinct interspecific and intraspecific vulnerability of coastal species to global change. Global Change Biol 27:3415–3431. https://doi.org/10.1111/gcb.15651
43. Nualart N, Ibáñez N, Soriano I, López-Pujol J (2017) Assessing the relevance of herbarium collections as tools for conservation biology. Bot Rev 83:303–325. https://doi.org/10.1007/s12229-017-9188-z

44. Orr MC, Hughes AC, Chesters D, Pickering J, Zhu CD, Ascher JS (2021) Global patterns and drivers of bee distribution. Curr Biol 31:451–458. https://doi.org/10.1016/j.cub.2020.10.053

45. Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT (2007) Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. J Biogeogr 34:102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

46. Powers RP, Jetz W (2019) Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios. Nat Clim Change 9:323–329. https://doi.org/10.1038/s41558-019-0406-z

47. Qin HN, Yang Y, Dong SY et al (2017) Threatened species list of China's higher plants. Biodivers Sci 2:696–744. https://doi.org/10.17520/biods.2017144

48. Raes N, ter Steege H (2007) A null-model for significance testing of presence-only species distribution models. Ecography 30:727–736. https://doi.org/10.1111/j.2007.0906-7590.05041.x

49. Richards DR, Thompson BS, Wijedasa L (2020) Quantifying net loss of global mangrove carbon stocks from 20 years of land cover change. Nat Commun 11:4260. https://doi.org/10.1038/s41467-020-18118-z

50. Román-Palacios C, Wiens JJ (2020) Recent responses to climate change reveal the drivers of species extinction and survival. Proc Natl Acad Sci USA 117:4211–4217. https://doi.org/10.1073/pnas.1913007117

51. Shi CY, Lai WF, Wen GW, Jiang TY, Zhu XR, Lv ZW, Zhang GF (2022) Prediction of potentially suitable area of *Fraxinus mandshurica* based on MaxEnt model. J Northwest Fores Univers 37:149–156. https://doi.org/10.3969/j.issn.1001-7461.2022.2.20

52. Sun L (2014) The genetic diversity and phylogeography of Litsea auriculata, an endemic and rare plant in China. Nanjing University, Nanjing

53. Tang ZY, Wang ZH, Zheng CY, Fang JY (2006) Biodiversity in China's mountains. Front Ecol Environ 4:347–352. https://doi.org/10.1890/1540-9295(2006)004[0347:Bicm]2.0.Co;2

54. Thiers BM (2020) The world's herbaria 2019: A summary report based on data from Index Herbariorum. New York Botanical Garden, USA

55. Wei XZ, Jiang MX (2012) Limited genetic impacts of habitat fragmentation in an old rare relict tree, *Euptelea pleiospermum* (Eupteleaceae). Plant Ecol 213:909–917. https://doi.org/10.1007/s11258-012-0052-2

56. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, NCEAS Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. Divers distrib 14:763–773. https://doi.org/10.1016/S0304-3800(01)00388-X

57. Wolmarans R, Robertson MP, van Rensburg BJ (2010) Predicting invasive alien plant distributions: how geographical bias in occurrence records influences model performance. J Biogeogr 37:1797–1810. https://doi.org/10.1111/j.1365-2699.2010.02325.x

58. Wu XK, Nan CH, Tang GG, Li Y, Mao LJ, Zhang ZC (2016) Impact of climate change on potential distribution range and spatial pattern of *Phoebe chekiangensis*. J Nanjing Fores Univers (Nat Sci Edition) 40:85–91. https://doi.org/10.3969/j.issn.1000-2006.06.013

59. Wu Y, Wang HF, Mu LQ (2022) Research progress and prospect of species distribution models. J Sci Teachers' Coll Univers 42:66–70. https://doi.org/10.3969/j.issn.1007-9837.2022.05.012

60. Xie D, Liu XQ, Chen YX et al (2021) Distribution and conservation of threatened gymnosperms in China. Global Ecol Conserv 32:e01915. https://doi.org/10.1016/j.gecco.2021.e01915

61. Yan Y, Li Y, Wang WJ et al (2017) Range shifts in response to climate change of *Ophiocordyceps sinensis*, a fungus endemic to the Tibetan Plateau. Biol Conserv 206:143–150.

https://doi.org/10.1016/j.biocon.2016.12.023

62. Yang Y, Lee BR, Heberling JM, Primark RB (2022) Herbarium specimens may provide biased flowering phenology estimates for dioecious species. Int J Plant Sci 183:777–783. https://doi.org/10.1086/722294
63. Yang YC, Huang PH (1982) Litsea Lam. Flora Reipublicae Popularis Sinicae Tomus 31. Science, Beijing
64. Zhang ZJ, Yan YJ, Tian Y, Li JS, He JS, Tang ZY (2015) Distribution and conservation of Orchid species richness in China. Biol Conserv 181:64–72. https://doi.org/10.1016/j.biocon.2014.10.026
65. Zheng WY, Zeng WH, Tang YS, Shi W, Cao KF (2018) Species diversity and biogeographical patterns of Lauraceae and Fagaceae in northern tropical and subtropical regions of China. Acta Ecol Sin 38:8676–8687. https://doi.org/10.5846/stxb201808281841
66. Zhou R, Ci XQ, Xiao JH, Cao GL, Li J (2021) Effects and conservation assessment of climate change on the dominant group-The genus *Cinnamomum* of subtropical evergreen broad-leaved forests. Biodivers Sci 29:697–711. https://doi.org/10.17520/biods.2020482
67. Zhu YY, Xu XT (2019) Effects of climate change on the distribution of wild population of *Metasequoia glyptostroboides*, an endangered and endemic species in China. Chin J Ecol 38:1629–1636. https://doi.org/10.13292/j.1000-4890.201906.018

Figures

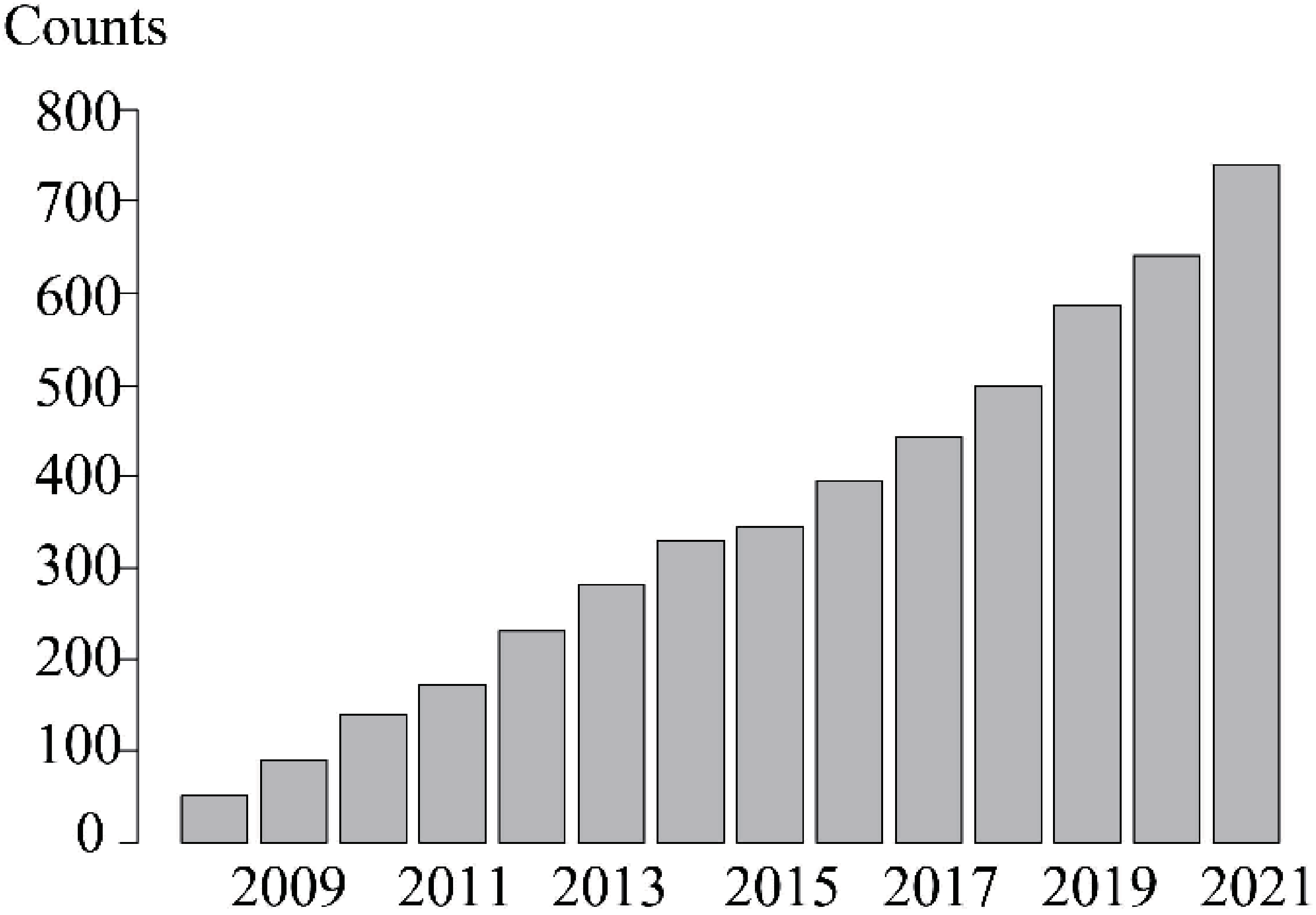

Figure 1

Web of Science changes in the literature of SDMs studies using MaxEnt in the database

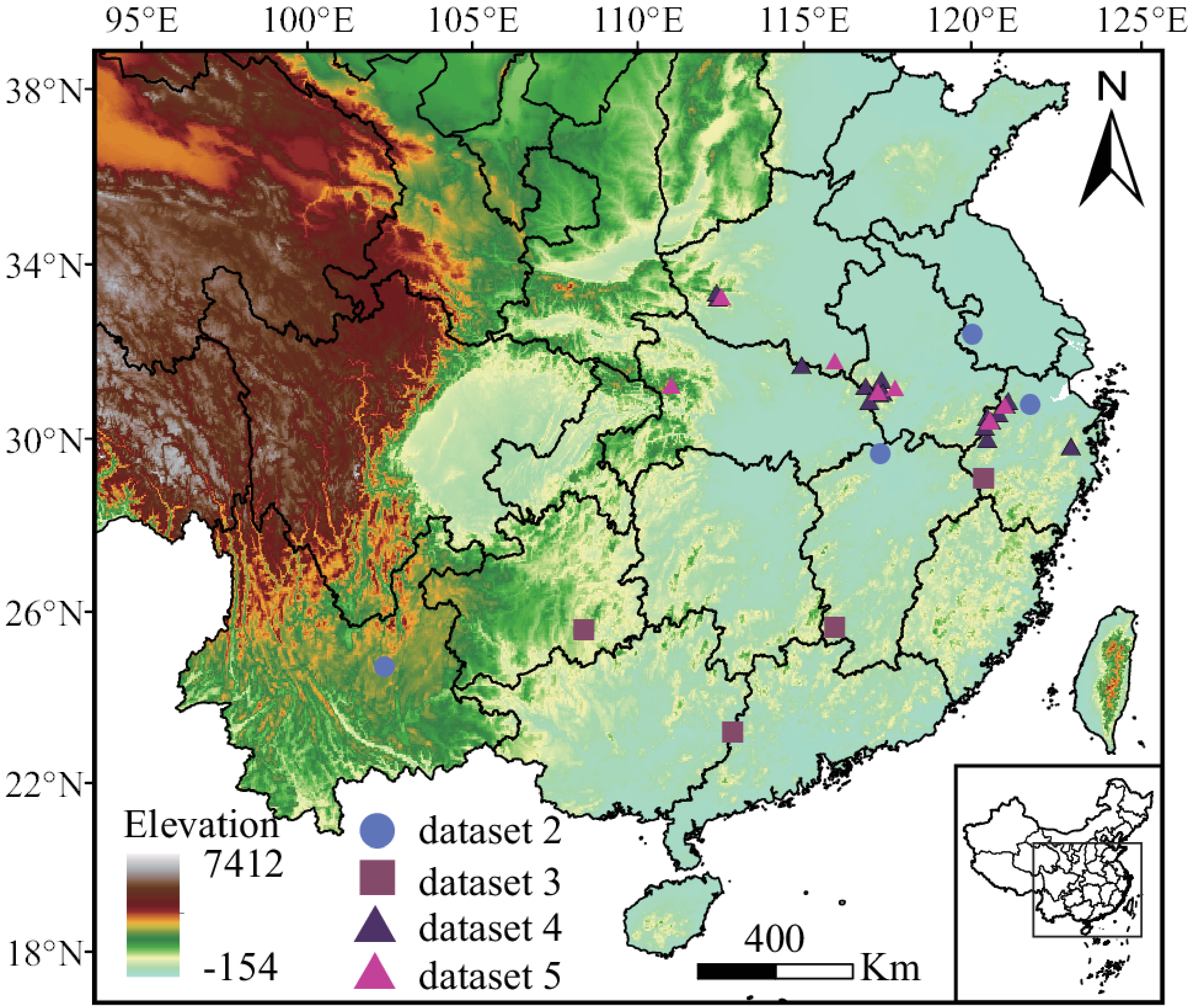

Figure 2

Distribution of *Litsea auriculata* according to different datasets (dataset 1 consisting of dataset 4 and dataset 5; dataset 6 including dataset 1, dataset 2, and dataset 3)

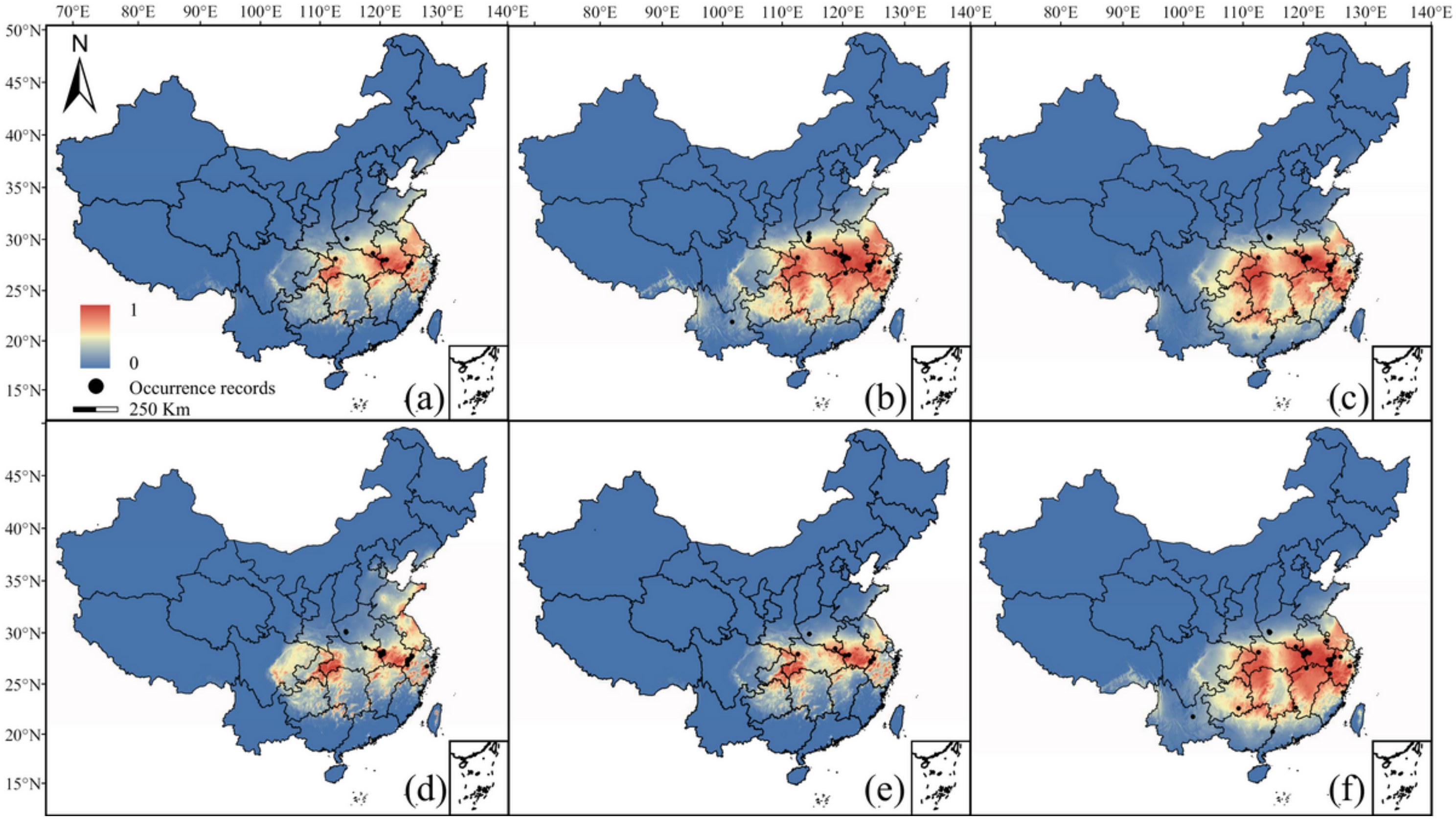

Figure 3

Potential distribution patterns of *Litsea auriculata* under current climatic conditions. (a) dataset 1; (b) dataset 2; (c) dataset 3; (d) dataset 4; (e) dataset 5; (f) dataset 6.

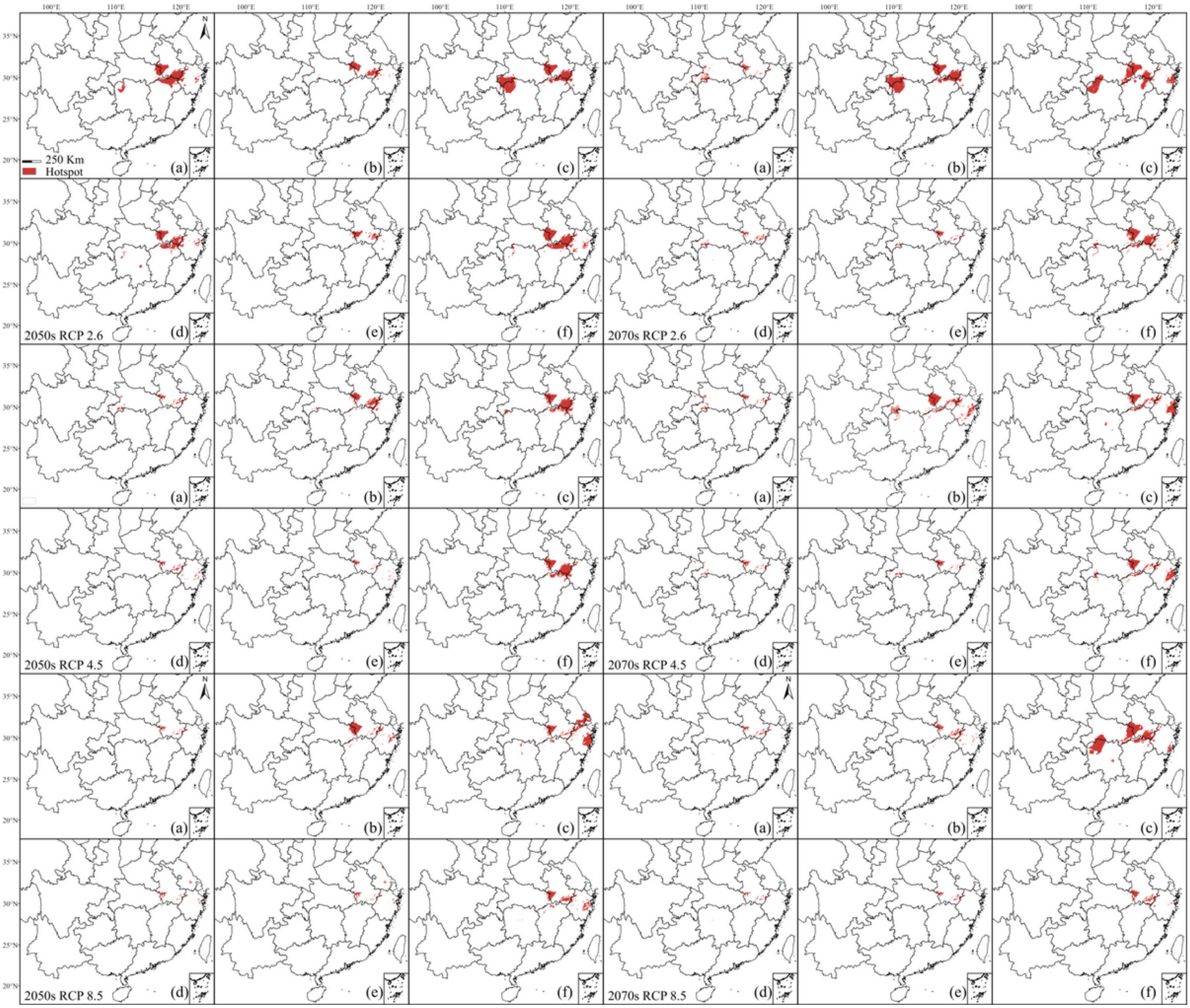

Figure 4

Projected hotspot regions of *Litsea auriculata* based upon different datasets under the 2050s and 2070s RCP 2.6/4.5/8.5 climate scenarios. (a) dataset 1; (b) dataset 2; (c) dataset 3; (d) dataset 4; (e) dataset 5; (f) dataset 6.

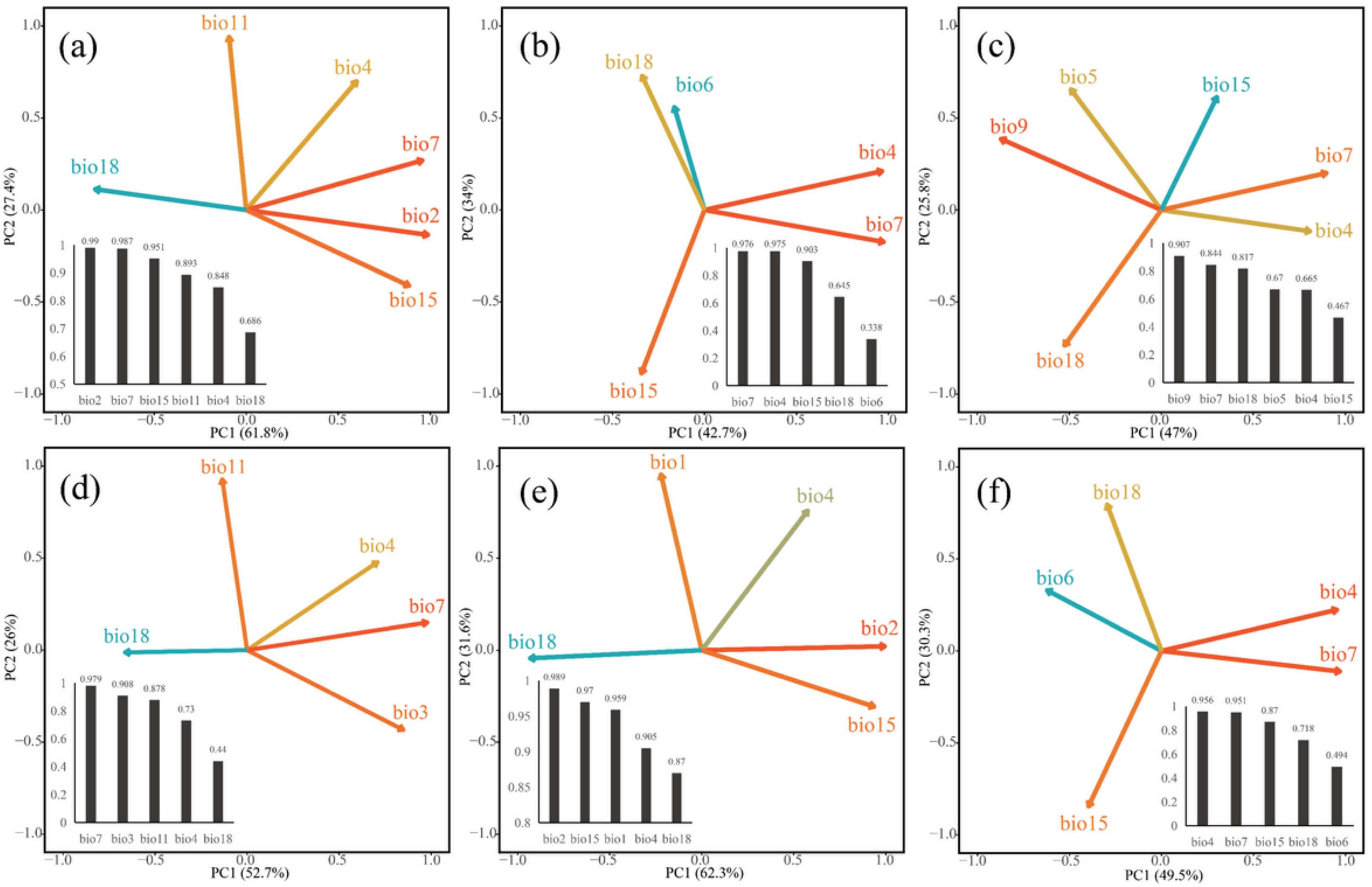

Figure 5

PCA of *Litsea auriculata* under current climatic condition. (a) dataset 1; (b) dataset 2; (c) dataset 3; (d) dataset 4; (e) dataset 5; (f) dataset 6. The bar chart represents the contribution of variables.

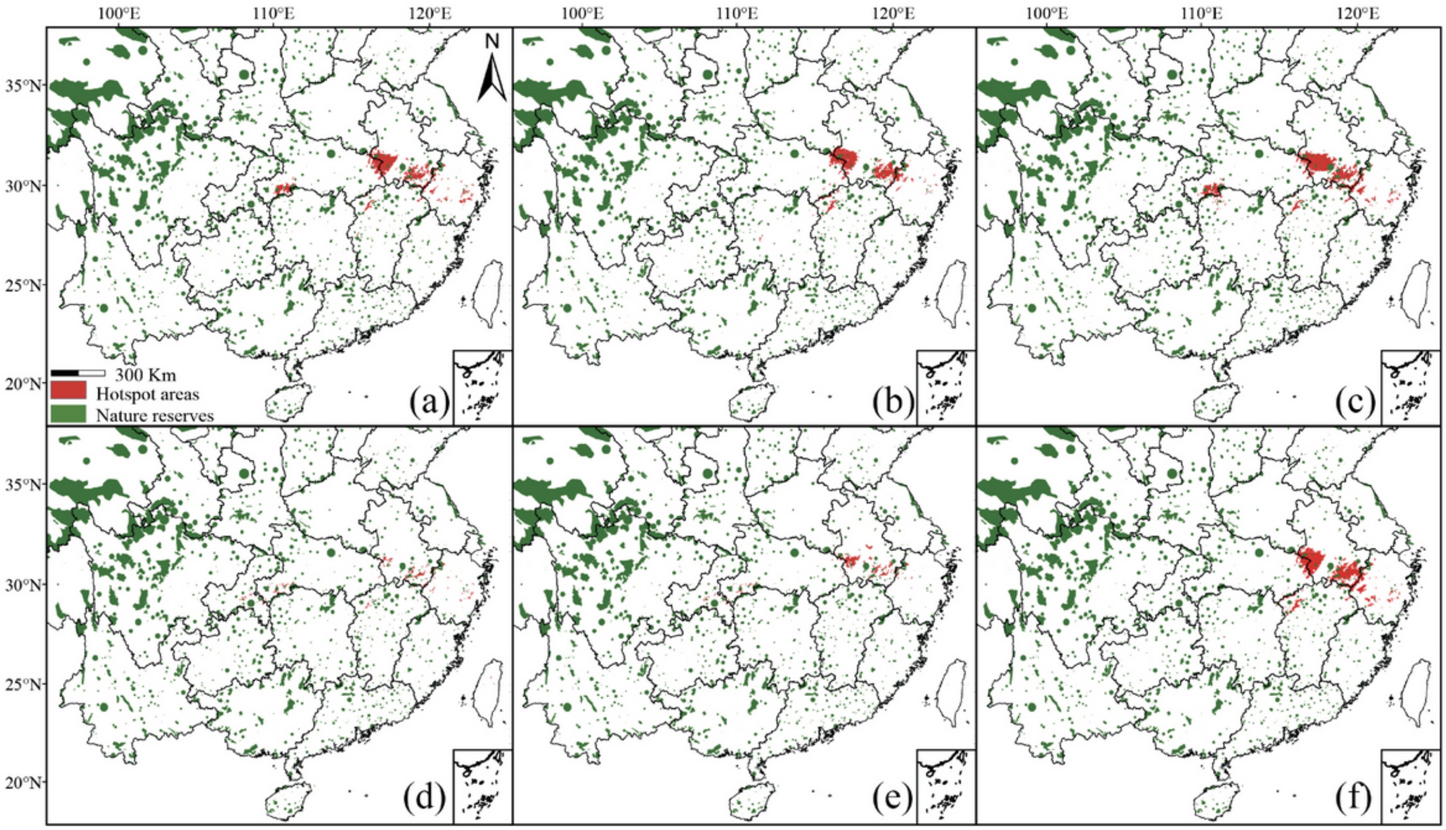

Figure 6

Hotspots located in all protected areas under current climate condition. (a) dataset 1; (b) dataset 2; (c) dataset 3; (d) dataset 4; (e) dataset 5; (f) dataset 6.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supportinginformation.docx