SCIENCE CHINA
Earth Sciences

•RESEARCH PAPER•

March 2022 Vol.65 No.3: 518–535
https://doi.org/10.1007/s11430-021-9867-1

Biome reconstruction on the Tibetan Plateau since the Last Glacial Maximum using a machine learning method

Feng QIN[1*], Yan ZHAO[1,3] & Xianyong CAO[2]

[1] *Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;*
[2] *Alpine Paleoecology and Human Adaptation (ALPHA) Group, State Key Laboratory of Tibetan Plateau Earth System Science (LATPES), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China;*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received May 7, 2021; revised October 29, 2021; accepted November 11, 2021; published online January 4, 2022

Abstract Historical biome changes on the Tibetan Plateau provide important information that improves our understanding of the alpine vegetation responses to climate changes. However, a comprehensively quantitative reconstruction of the historical Tibetan Plateau biomes is not possible due to the lack of quantitative methods that enable appropriate classification of alpine biomes based on proxy data such as fossil pollen records. In this study, a pollen-based biome classification model was developed by applying a random forest algorithm (a supervised machine learning method) based on modern pollen assemblages on and around the Tibetan Plateau, and its robustness was assessed by comparing its results with the predictions of the biomisation method. The results indicated that modern biome distributions reconstructed using the random forest model based on modern pollen data generally concurred with the observed zonal vegetation. The random forest model had a significantly higher accuracy than the biomisation method, indicating the former is a more suitable tool for reconstructing alpine biome changes on the Tibetan Plateau. The random forest model was then applied to reconstruct the Tibetan Plateau biome changes from 22 ka BP to the present based on 51 fossil pollen records. The reconstructed biome distribution changes on the Tibetan Plateau generally corresponded to global climate changes and Asian monsoon variations. In the Last Glacial Maximum, the Tibetan Plateau was mainly desert with subtropical forests distributed in the southeast. During the last deglaciation, the alpine steppe began expanding and gradually became zonal vegetation in the central and eastern regions. Alpine meadow occupied the eastern and southeastern areas of the Tibetan Plateau since the early Holocene, and the forest-meadow-steppe-desert pattern running southeast to northwest on the Tibetan Plateau was established afterwards. In the mid-Holocene, subtropical forests extended north, which reflected the "optimum" condition. During the late Holocene, alpine meadows and alpine steppes expanded south.

Keywords Biome reconstruction, Random forest algorithm, Biomisation method, Pollen data, Last Glacial Maximum, Tibetan Plateau

Citation: Qin F, Zhao Y, Cao X. 2022. Biome reconstruction on the Tibetan Plateau since the Last Glacial Maximum using a machine learning method. Science China Earth Sciences, 65(3): 518–535, https://doi.org/10.1007/s11430-021-9867-1

1. Introduction

The uplift of the Tibetan Plateau altered the intensity of the Asian monsoon (An et al., 2001, 2015; Molnar et al., 2010) and formed a unique alpine ecosystem (Zhang, 1978; Zheng et al., 1979; Wu, 1980). The development and response of alpine vegetation on the Tibetan Plateau due to climate changes have garnered considerable attention (e.g., Tang and Li, 2001; Chen et al., 2020; Zhao et al., 2020). The biome dynamics since the Last Glacial Maximum (LGM) represent

* Corresponding author (email: qinfeng@igsnrr.ac.cn)

a series of vegetation responses to climate changes from glacial to interglacial, and biome changes during this period have been studied based on fossil pollen sequences at many fossil sites across the Tibetan Plateau (e.g., Shen et al., 2006; Zhao et al., 2011; Herzschuh et al., 2014; Li et al., 2019; Shi et al., 2020).

Several reviews have provided qualitative overviews of the zonal vegetation shifts in different parts of the Tibetan Plateau since the LGM. For instance, Tang and Li (2001) summarised the fossil pollen data and described the temporal-spatial biome distributions on the Tibetan Plateau during the Holocene; Tang et al. (2021) reviewed the published palaeopalynological works of the Tibetan Plateau, and illustrated the major biome changes in different sites during 20–0 ka BP.

Quantitative reconstruction of past biome changes based on the fossil pollen data on the Tibetan Plateau is rare, although fossil pollen records are available at many sites. All the quantitative biome reconstructions were performed by applying the biomisation method (Prentice et al., 1996). Herzschuh et al. (2006) presented the Holocene pollen record of Zigetang Co, and quantitatively revealed the local biome variations between temperate steppe and alpine steppe during the Holocene with a site-specific biome classification scheme. Herzschuh et al. (2009) reconstructed the changes of tundra and steppe biomes surrounding Koucha Lake since the late glacial. Dallmeyer et al. (2011) reconstructed the biome changes over the past 6 ka at four sites representing different vegetation zones on the Tibetan Plateau, and exhibited the variations of the forest, shrub, steppe/meadow, and desert at these sites. Despite applying the same method, these pollen-based biome reconstructions used different biome classification schemes, so their results are not fully comparable.

Minimal studies on past biome distributions in China included quantitative reconstructions of past biomes on the Tibetan Plateau (Yu et al., 2000; Ni et al., 2014; Sun et al., 2020) using fossil pollen data and biomisation method (Prentice et al., 1996). These studies aimed to determine past biome distributions across China at a subcontinental scale; however, insufficient attention was given to specific regions such as the Tibetan Plateau. The biome classification scheme of Yu et al. (2000) only incorporated one tundra biome type to represent alpine vegetation of the Tibetan Plateau and presented the past biome distributions for two time windows (18 and 6 ka BP). Ni et al. (2014) reconstructed the biome changes along a continuous timeline from the LGM, and their biome scheme contained multiple tundra biome types; however, the tundra types did not distinctly conform to the zonal vegetation of the Tibetan Plateau. Sun et al. (2020) reconstructed the biomes of China in the mid-Holocene, and their scheme included two alpine biome types compatible with the vegetation zones of the Tibetan Plateau. However, a suitable quantitative reconstruction method for the biome changes on the Tibetan Plateau since the LGM has yet to be established.

Two studies performed model simulations for past biome distribution on the Tibetan Plateau. Song et al. (2005) simulated the early Holocene biome distribution of the Tibetan Plateau applying a biogeography-biogeochemistry model (BIOME4), and the modeled palaeo-biome pattern was validated using a biome zonation map inferred from fossil pollen data qualitatively. Dallmeyer et al. (2011) simulated the biome changes on the Tibetan Plateau over the past 6 ka using a coupled atmosphere-ocean-vegetation model (ECHAM5/JSBACH-MPIOM), and assessed the model performance by using the biome reconstructions from fossil pollen data at four representative sites. However, the pollen-based biome reconstructions used in both cases didn't provide sufficient evidence for verifying the model results because of using either qualitative interpretation at a coarse spatial scale or quantitative reconstructions from a limited number of sites. A more comprehensive pollen-based reconstruction of past biome pattern on the Tibetan Plateau applying a suitable quantitative method will facilitate the model-proxy comparisons for identifying the mechanism of past biome changes.

The random forest algorithm, a supervised machine learning method, was recently used to establish a pollen-biome model to reconstruct historical biomes. Using modern pollen assemblages from Africa and the Arabian Peninsula, Sobol and Finkelstein (2018) tested the reliability of eight numerical biome prediction methods based on pollen data and determined that the random forest model most accurately predicted biomes from pollen data. Subsequently, Sobol et al. (2019) applied the random forest algorithm to develop pollen-biome classification models in Southern Africa, resulting in highly accurate modern biome predictions based on modern pollen data, and they successfully used the random forest models to reconstruct biome changes in the last 60 ka at Wonderkrater.

The machine learning method has great potential for biome reconstruction of the Tibetan Plateau; however, validation of its robustness is necessary. In this study, we introduced the random forest algorithm to the quantitative reconstruction of past biome changes on the Tibetan Plateau. A pollen-biome reconstruction model for the Tibetan Plateau was developed applying the random forest algorithm based on a modern pollen dataset. The robustness of the proposed model was assessed by comparing its prediction accuracy with the results of the biomisation method using the same modern pollen dataset. Finally, the biome changes at fossil sites across the Tibetan Plateau were quantitatively reconstructed to illustrate the successive vegetation changes on the Tibetan Plateau since the LGM.

2. Materials and methods

2.1 Study region

The Tibetan Plateau covers ca. 2.54 million km^2 (Zhang et al., 2014) with an average elevation of >4000 m a.s.l. Its climate is mainly controlled by the Indian summer monsoon, the East Asian summer monsoon, and the mid-latitude westerlies (Chen et al., 2020). The vegetation zonation on the plateau along the thermal and moisture gradients from southeast to northwest is generally forest-meadow-steppe-desert (Zhang, 1978, 2007; Zheng et al., 1979).

The tropical rain forests are restricted to low-altitude areas on the southern slope of the Himalayas and are dominated by semi- and tropical evergreen trees such as *Dipterocarpus* spp., *Artocarpus chaplasha*, *Dysoxylum* spp., *Canarium resiniferum*, *Tetrameles nudiflora*, *Altingia excelsa*, *Chukrasia tabularis*, and *Shorea assamica*. Subtropical broadleaf evergreen forests occur in the southeastern portion of the Tibetan Plateau, and the dominant species mainly include evergreen trees in the Fagaceae family, such as *Castanopsis* spp., *Lithocarpus* spp., *Cyclobalanopsis* spp., evergreen Quercus (*Q. aquifolioides*, *Q. rehderiana*, and *Q. semicarpifolia*), and other evergreen genera such as *Schima*, *Machilus*, *Manglietia*, and *Ficus*. Coniferous trees, such as *Pinus densata*, *P. yunnanesis*, and *Tsuga dumosa*, play important roles in some of these forests. The high-altitude mountain area can be occupied by cold temperate needleleaf forests, with dominant species including *Picea* (*P. likiangensis*, *P. asperata*, *P. aurantiaca*, *P. purpurea*, and *P. brachytyla* var. *complanata*), *Abies* (*A. georgei* and *A. squamata*), *Sabina* (*S. tibetica* and *S. saltuaria*), *Pinus* (*P. griffithii*), and *Larix* (*L. chinensis*) (Zhang, 2007).

The eastern Tibetan Plateau is characterised by subalpine scrubs and alpine meadows. Dominant shrubs include *Rododendron* (*Rh. capitatum*, *Rh. thimifolium*, *Rh. przewalskii*, *Rh. violaceum*, *Rh. litangensis*, *Rh. nivale*, *Rh. cephalanthoides*, and *Rh. fastigiatum*), *Salix* (*S. cupularis*, *S. oriterpha*, *S. atopantha*, and *S. sclerophylla*), *Potentilla fruticosa*, *Rosa sericea*, *Sibiraea angustata*, *Spiraea alpina*, and *Caragana jubata*. The alpine meadows are characterised by sedges such as *Kobresia* (*K. pygmaea*, *K. humilis*, *K. setchwanensis*, *K. capillifolia*, and *K. prattii*) and *Carex* (*C. lanceolata*, *C. muliensis*, and *C. meyeriana*), and grass species such as *Elymus nutans*, *Roegneria nutans*, *Stipa purpurea*, *S. aliena*, and *Deschampsia caespitosa*. Diverse forbs in genera such as *Polygonum*, *Potentilla*, *Anaphalis*, *Leontopodium*, *Taraxacum*, *Saussurea*, *Pedicularis*, *Anemone*, *Trollius*, *Ranunculus*, *Thalictrum*, *Gentiana*, *Swertia*, *Oxytropis*, *Astragalus* are also frequently found in the alpine meadows (Zhang, 2007).

Alpine steppes prevail in the central portion of the Tibetan Plateau, which are mainly dominated by plants from the genera of *Stipa* (*S. purpurea*, *S. bungeana*, *S. subsessiliflora* var. *basiplumosa*, *S. roborowskyii*, and *S. capillacea*), *Artemisia* (*A. wellbyi*, *A. younghusbandii*, and *A. stracheyi*) and *Carex* (*C. moorcroftii* and *C. montis-everestii*). Other important components include grasses such as *Littledalea racemose*, *Orinus thoroldii*, *Pennisetum flaccidum*, and *Aristida adscensionis*, and shrubs such as *Caragana versicolor*, *Sophora moorcroftiana*, *Sabina pingii* var. *wilsonii*, and *Potentilla fruticosa* (Zhang, 2007).

Alpine and temperate deserts occupy the northern and western regions of the Tibetan Plateau. The dominant species are mainly plants from the Chenopodiaceae family, such as *Ceratoides* (*C. compacta* and *C. latens*), *Salsola* (*S. abrotanoides*), *Haloxylon* (*H. ammodendron*), and *Kalidium* (*K. foliatum* and *K. cuspidatum*), and xerophilous plants such as *Ephedra* (*E. przewalskii* and *E. intermedia*), *Zygophyllum xanthoxylon*, *Nitraria* (*N. roborowskii* and *N. sibirica*), *Ajania tibetica* and *Artemisia* (*A. rhodantha* and *A. arenaria*) can also be dominant. Some grasses such as *Stipa purpurea* and *S. glareosa*, and sedges like *Carex moorcroftii*, together with the above-mentioned typical desert plants, form the desert-steppe vegetation (Zhang, 2007).

2.2 Modern pollen dataset

The modern pollen dataset used as the training set for developing the pollen-biome classification model contained 1802 samples in 17 vegetation zones obtained from locations on and surrounding the Tibetan Plateau (Cao et al., 2014; Zheng et al., 2014; Zhao et al., 2021) (Figure 1). Modern pollen samples covered a significantly larger area than the extent of the Tibetan Plateau to incorporate more vegetation types represented by the fossil pollen records. The modern pollen assemblages in the dataset were from surface soils (832 samples), moss polsters (619 samples), and surface lake sediments (351 samples). In the study region, pollen assemblages from a single sediment type cannot cover all target vegetation zones (Figure 1). Therefore, we combined the pollen assemblages from different types of sediment to form the modern pollen dataset, although pollen signals may differ among sediment types (Fall, 1992; Wilmshurt and McGlone, 2005; Zhao et al., 2009; Lisitsyna et al., 2012). The dataset contained 504 terrestrial pollen taxa, which were homogenised into 230 taxa by merging the synonyms and combining the low-level taxonomic groups. The pollen percentages were recalculated based on the sum of the terrestrial pollen.

A large number of the modern pollen samples had no original vegetation information in the dataset, and the original vegetation data of the other samples were probably obtained via different criteria such as different field survey extents and vegetation classification systems. Therefore, the biome labels for all modern pollen assemblages were reassigned according to their positions on the vegetation map to

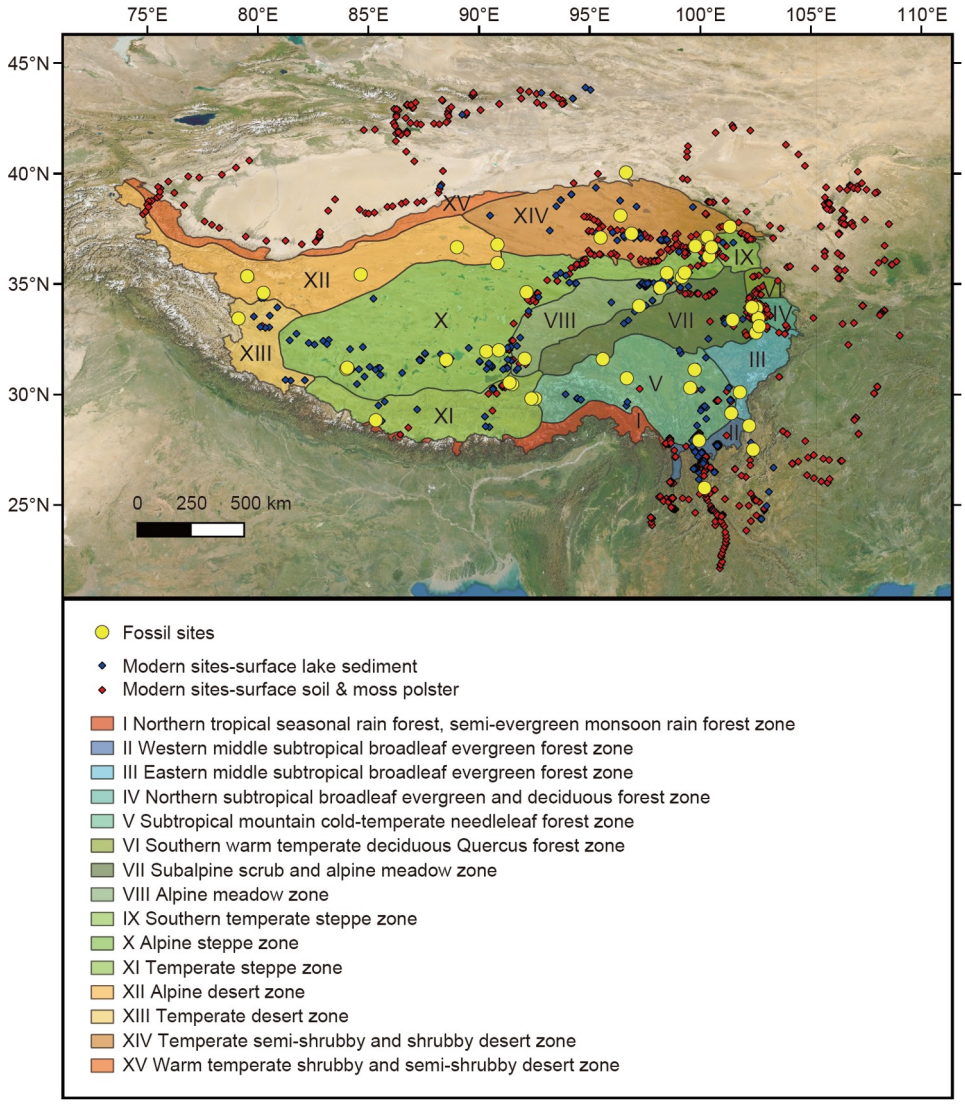

Figure 1 Map showing the vegetation zonation of the Tibetan Plateau and the fossil and modern pollen sample locations.

ensure consistency, so each pollen assemblage was assigned as a biome type that represented the relevant vegetation zone. The vegetation zone is the subordinate regionalisation unit to the vegetation region in the vegetation regionalisation map of China (Zhang, 2007). Such a spatial scale should be appropriate for establishing pollen-vegetation model, since studies on modern pollen-vegetation relationships indicated that modern pollen assemblages can reflect vegetation zones better than finer-scale vegetation information (Felde et al., 2014; Qin, 2021).

2.3 Fossil pollen dataset

A total of 51 fossil pollen records from the Tibetan Plateau (Figure 1, Appendix Table A1, https://link.springer.com) were used to reconstruct past biome changes from the LGM to the present, which were primarily obtained from the Late Quaternary pollen dataset of eastern continental Asia (Cao et al., 2013). The elevations of the fossil sites ranging from 1974 to 5325 m a.s.l. Twelve sites were located in the modern subtropical forest region, which included four and eight sites in the middle subtropical broadleaf evergreen and subtropical mountain cold-temperate needleleaf forest zones, respectively. Seven sites were in the subalpine scrub and alpine meadow zone, and three were in the alpine meadow zone. Fifteen sites were in the alpine (11 sites), temperate (3 sites), and southern temperate (1 site) steppe zones and the remaining 14 sites were in desert zones, including the alpine (5 sites), temperate (1 site), and temperate semi-shrubby and shrubby (8 sites) desert zones. The original 269 fossil pollen taxa were homogenised into 135 taxa, and all the fossil pollen taxa were harmonised to correspond with the modern pollen dataset.

Most of the fossil pollen records had at least three age

control points. New age-depth models were developed for most of the records using the Bayesian method (Blaauw and Christen, 2011) and constructed using rbacon 2.3.7 (Blaauw and Christen, 2019) in R v. 3.5.1 (R Core Team, 2018). The original age-depth models of lakes Donggi Cona, Muge Co, Taro Co, Ximen Co, and Wuxu Lake were developed using the Bayesian method, and the age-depth model of Dunde ice cap was determined by the annual layers of the ice core. These original age-depth models were used in this study. The sample resolutions of 36 records were greater than 500 years/sample, 12 records had sample resolutions between 500–1000 years/sample, and 3 records had sample resolutions >1000 years/sample (Appendix Table A1). All fossil pollen records were linearly interpolated to 500-year time slices, resulting in a studied time interval of 22–0 ka BP containing 45 time windows. It's noteworthy that some fossil pollen records had limited age control points, or showed low sampling resolutions (Appendix Table A1). These fossil pollen records were all included in the palaeobiome reconstructions to maximise the number of fossil sites and increase the spatial representation, although uncertainties may be involved in the ages of these records. Nevertheless, these uncertainties may be reduced to some extent by focusing on the long-term changes of large-scale biome patterns revealed by multiple sites rather than on the biome succession of a single site.

2.4 Biome reconstruction methods

2.4.1 Random forest algorithm

Modern pollen assemblages and their vegetation type information were used to train the pollen-biome classification model via the random forest algorithm (Breiman, 2001). Random forest algorithm fits many classification trees, and the final prediction is made according to the majority vote from all the trees (Breiman, 2001; Cutler et al., 2007). Each tree is fitted to a random bootstrap subset of the original dataset using classification and regression tree (CART) methodology (Breiman et al., 1984). At each node of the classification tree, only a small group of variables are randomly selected to split on. After being fully grown (not pruned), each tree is tested on the out-of-bag (OOB) samples that are not included in the bootstrap subset (about one-third of the original dataset), and the error rates are estimated. The averaged error rates over all trees (OOB error rates) are used to assess the performance of the random forest model. The importance of a specific variable (pollen taxon in this study) is evaluated by using the mean decrease in accuracy, which measures the increase in misclassification when the values of the variable are randomly permuted for the OOB samples. The model training was performed following the procedures of Sobol and Finkelstein (2018) and Sobol et al. (2019). The number of trees and pollen taxa randomly selected as predictor variables at each node were set to 500 and 11, respectively. Biome types containing <15 samples were excluded from model fitting. The algorithm was repeated 100 times, and the model with the lowest OOB error rate was selected for subsequent analyses. The pollen-biome model was applied to the fossil pollen data from the Tibetan Plateau to reconstruct the biome changes for each fossil site. The reconstructed biomes were mapped, and the biome distributions on the Tibetan Plateau since the LGM in the different time windows were determined. The establishment and application of the random forest model were implemented in R v.3.5.1 (R Core Team, 2018) using the randomForest 4.6–14 package (Liaw and Wiener, 2002).

2.4.2 Biomisation method

The biomisation method (Prentice et al., 1996) was also employed to reconstruct modern biomes from modern pollen data for comparison with the resultant random forest model biomes. The process included (Prentice et al., 1996; Prentice and Webb III, 1998): (1) assigning each pollen taxon to one or more plant function types (PFTs) according to its ecophysiological and bioclimatic characteristics; (2) defining biomes based on the characteristic PFTs; (3) constructing a biome × taxon matrix indicating which pollen taxa might occur in each biome; and (4) calculating the affinity scores for each biome using pollen percentage data as follows:

$$A_{ik} = \sum_j \delta_{ij} \sqrt{\left\{\max\left[0, \left(p_{jk} - \theta_j\right)\right]\right\}},$$

where A_{ik} is the affinity score of pollen sample k for biome i, δ_{ij} is the entry in the biome × taxon matrix for biome i and taxon j, p_{jk} is the pollen percentage for taxon j in sample k, and θ_j is the threshold pollen percentage (0.5% in this study). Finally, the pollen assemblage was assigned to the biome with the highest affinity score. When affinity scores of different biomes are equal, the biome defined by a smaller number of PFTs has priority (Prentice et al., 1996). The classification of PFTs and biomes was based on the Sun et al. (2020) scheme (detailed classifications see Tables B1–B3 in Appendix B) because it included more alpine biome types that correspond to the alpine vegetation of the Tibetan Plateau and more accurately reconstructed the modern biomes of China than previous studies (Sun et al., 2020). The biomisation method was performed using 3Pbase software (Guiot and Goeury, 1996).

3. Results

3.1 Random forest model performance

The random forest model was established based on 1764 samples from 13 biomes (>15 samples for each biome), including northern tropical seasonal and semi-evergreen

monsoon rain forest, southern subtropical monsoon broadleaf evergreen forest, middle subtropical broadleaf evergreen forest, northern subtropical broadleaf evergreen and deciduous forest, southern warm-temperate deciduous *Quercus* forest, subtropical mountain cold-temperate needleleaf forest, subalpine scrub & alpine meadow, alpine meadow, southern temperate steppe, alpine steppe, temperate steppe, warm-temperate shrubby and semi-shrubby desert, and temperate semi-shrubby and shrubby desert.

The overall performance of the random forest model was ideal with an overall OOB error rate of 23.47%. Reconstructed biomes produced by the random forest model generally showed a zonal distribution similar to the modern vegetation regionalisation (Figure 2). The model accuracies for the different biome types varied and are presented in Table 1.

The model showed the highest classifying accuracy for the middle subtropical broadleaf evergreen forest biome (93.56%), and most of the misidentified samples were assigned to the alpine steppe (2.99%) and subtropical mountain cold-temperate needleleaf forest (1.95%) biomes. The subtropical mountain cold-temperate needleleaf forest biome also exhibited a high accuracy (72.94%), and the majority of the misidentified samples were assigned to the middle subtropical broadleaf evergreen forest biome (23.53%).

The random forest model showed weak performances for the other forest biomes. The model correctly assigned 39.29% of the northern tropical seasonal and semi-evergreen monsoon rain forest samples, and misidentifications were primarily assigned to the middle subtropical broadleaf evergreen forest biome (57.14%). A large proportion (68.75%) of southern subtropical monsoon broadleaf evergreen forest samples was incorrectly assigned to the middle subtropical broadleaf evergreen forest biome, and only 29.69% was correctly assigned. The majority of the northern subtropical broadleaf evergreen and deciduous forest samples were assigned to the subalpine scrub & alpine meadow (38.46%), middle subtropical broadleaf evergreen forest (28.85%), and northern subtropical broadleaf evergreen and deciduous forest (19.23%) biomes. The southern warm-temperate deciduous *Quercus* forest biome showed the lowest accuracy (8.33%) among the forest biomes, which were inaccurately assigned to the northern subtropical broadleaf evergreen and deciduous forest (25.00%), middle subtropical broadleaf evergreen forest (20.83%), subalpine scrub & alpine meadow (20.83%), and southern temperate steppe (16.67%) biomes.

Most samples from the two desert biomes, warm-temperate shrubby and semi-shrubby desert (90.08%) and temperate semi-shrubby and shrubby desert (87.50%), were correctly assigned, and incorrect assignments for samples of the former were all assigned to the latter, while those of the latter were primarily assigned to the former.

The majority of the alpine meadow samples were correctly assigned (83.10%), and the remainder were mainly assigned as the alpine steppe biome (9.86%). The subalpine scrub & alpine meadow samples exhibited an accuracy of 80.43%, while 7.61% and 6.52% were incorrectly assigned to the alpine meadow and northern subtropical broadleaf evergreen and deciduous forest biomes, respectively.

The southern temperate steppe samples were primarily assigned correctly (82.5%), and most of the misidentified

Table 1 Confusion matrix of the random forest model performance for classifying biomes using modern pollen assemblages[a)]

Observed vs. predicted	NTRF	SStBEF	MStBEF	NStBEDF	StMNF	SWTDQF	SaScM	AM	STS	TS	AS	WTD	TD	Accuracy
NTRF	11	1	16	0	0	0	0	0	0	0	0	0	0	0.39
SStBEF	0	19	44	0	1	0	0	0	0	0	0	0	0	0.30
MStBEF	0	2	625	1	13	0	4	0	0	3	20	0	0	0.94
NStBEDF	0	0	15	10	1	3	20	0	2	0	0	0	1	0.19
StMNF	0	1	40	0	124	0	4	0	1	0	0	0	0	0.73
SWTDQF	0	0	5	6	0	2	5	0	4	0	1	0	1	0.08
SaScM	0	0	1	6	0	0	74	7	1	0	2	0	1	0.80
AM	0	0	0	0	0	0	2	59	1	0	7	0	2	0.83
STS	0	0	1	0	0	0	1	1	99	0	3	4	11	0.82
TS	0	0	7	1	0	0	1	1	5	6	13	0	2	0.17
AS	0	0	20	0	0	0	4	17	4	0	42	2	35	0.34
WTD	0	0	0	0	0	0	0	0	0	0	0	118	13	0.90
TD	0	0	0	0	0	0	1	0	6	0	5	11	161	0.87

a) NTRF, northern tropical seasonal and semi-evergreen monsoon rain forest; SStBEF, southern subtropical monsoon broadleaf evergreen forest; MStBEF, middle subtropical broadleaf evergreen forest; NStBEDF, northern subtropical broadleaf evergreen and deciduous forest; StMNF, subtropical mountain cold-temperate needleleaf forest; SWTDQF, southern warm-temperate deciduous *Quercus* forest; SaScM, subalpine scrub & alpine meadow; AM, alpine meadow; STS, southern temperate steppe; TS, temperate steppe; AS, alpine steppe; WTD, warm-temperate shrubby and semi-shrubby desert; TD, temperate semi-shrubby and shrubby desert

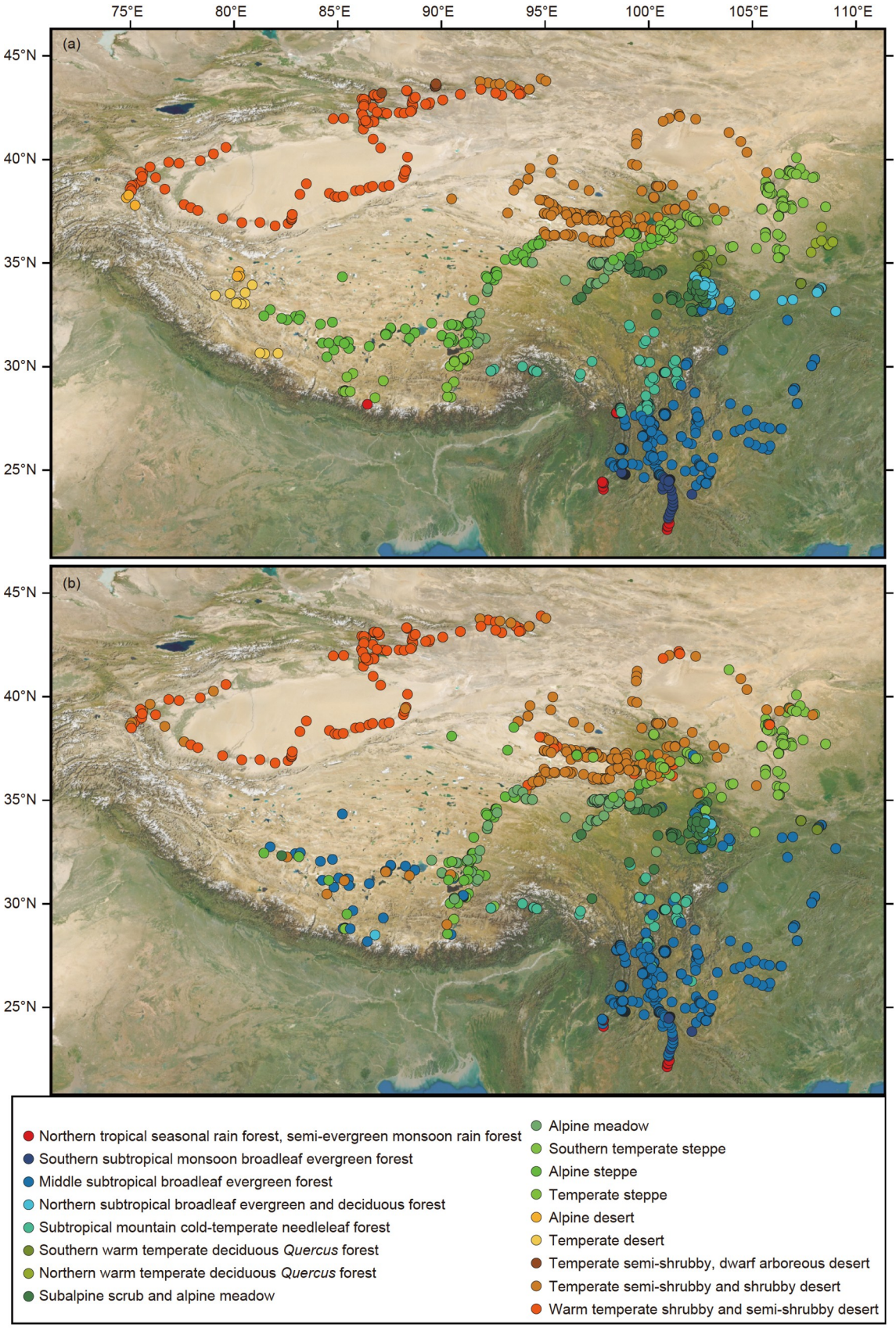

Figure 2 The observed modern biomes (a) and the predicted biomes produced via the random forest model (b) for the modern pollen samples.

samples were assigned into the temperate semi-shrubby and shrubby desert (9.17%). Samples from the other two steppe biomes showed low accuracy, with only 33.87% of the alpine steppe samples being correctly assigned, and the remainder were incorrectly assigned to the temperate semi-shrubby and shrubby desert (28.23%), middle subtropical broadleaf evergreen forest (16.13%), and alpine meadow (13.71%) biomes. Only 16.67% of the temperate steppe samples were correctly assigned, and 36.11%, 19.44%, and 13.89% were incorrectly assigned to the alpine steppe, middle subtropical broadleaf evergreen forest, and southern temperate steppe biomes, respectively.

The pollen taxa exhibited different effects on the accuracy of the model (Figure 3), and their importance for the model was calculated through the mean decrease in accuracy. *Pinus*, Amaranthaceae/Chenopodiaceae, and *Alnus* accounted for more than 5% of the mean decrease in accuracy, respectively, and played the most important role in accurately classifying the different vegetation zones. Cyperaceae, *Ephdera*, *Quercus* (Evergreen), *Artemisia*, *Betula*, and *Castanopsis*/*Lithocarpus* contributed significantly to an individual mean decrease in accuracy exceeding 2%, and the mean decrease in accuracy of *Tsuga*, *Abies*/*Picea*, *Nitraria*, Ranunculaceae, *Quercus* (Deciduous), *Juglans*, Asteraceae, and Ericaceae exceeded 1%.

3.2 Modern vegetation reconstructed via biomisation method

Fourteen biome types were reconstructed for modern pollen samples from the Tibetan Plateau and its vicinity using the biomisation method (Table 2, Figure 4, Appendix C), which included the tropical rain forest (TRFO), tropical seasonal forest (TSFO), south subtropical broadleaf evergreen forest (STFO), middle subtropical broadleaf evergreen forest (MTFO), north subtropical mixed forest (WAMF), warm-temperate mixed forest (TEDE), cool-temperate mixed forest (COMX), cold-temperate evergreen conifer forest (CLEC), cold-temperate summergreen conifer forest (CLDC), alpine meadow (ALME), alpine steppe (ALST), cool-temperate steppe (STEP), cool-temperate desert steppe (TEDS), and desert (DESE). In order to distinguish from the random forest model, the acronyms were used for biome names of biomisation method.

Based on a comparison with the observed biomes, only a few biome types reconstructed from modern pollen data showed meaningful accuracy (Table 2). The STEP samples showed the strongest agreement between the reconstructed and observed biomes, with 70.00% being correctly assigned, while 11.67% and 9.17% of the samples were inaccurately assigned to the COMX and DESE, respectively. Of the DESE samples 60.29% were correctly reconstructed; however, 29.86% were incorrectly assigned to the STEP. In addition to the STEP and DESE, the ALME and TSFO samples showed relatively high accuracies. For the ALME samples, 38.65% were correctly reconstructed, and 29.45% and 28.22% were incorrectly assigned to the ALST and STEP, respectively. The TSFO samples yielded an accuracy of 35.71%; however, 25.00% and 21.43% of the samples were

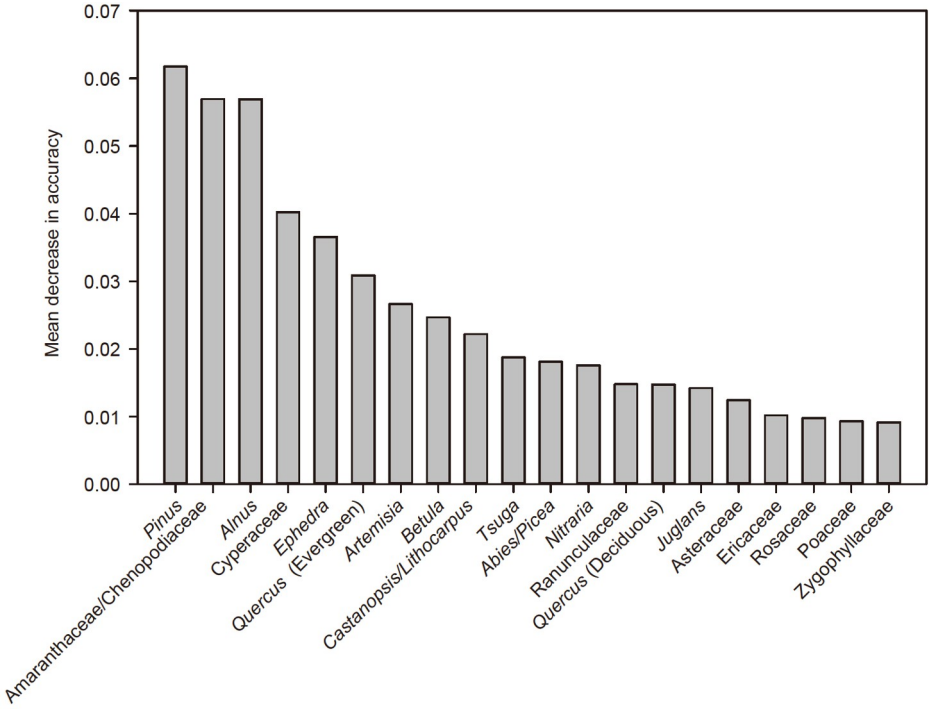

Figure 3 Pollen taxa significance determined by the mean decrease in accuracy in the pollen-based biome reconstruction model established using the random forest algorithm.

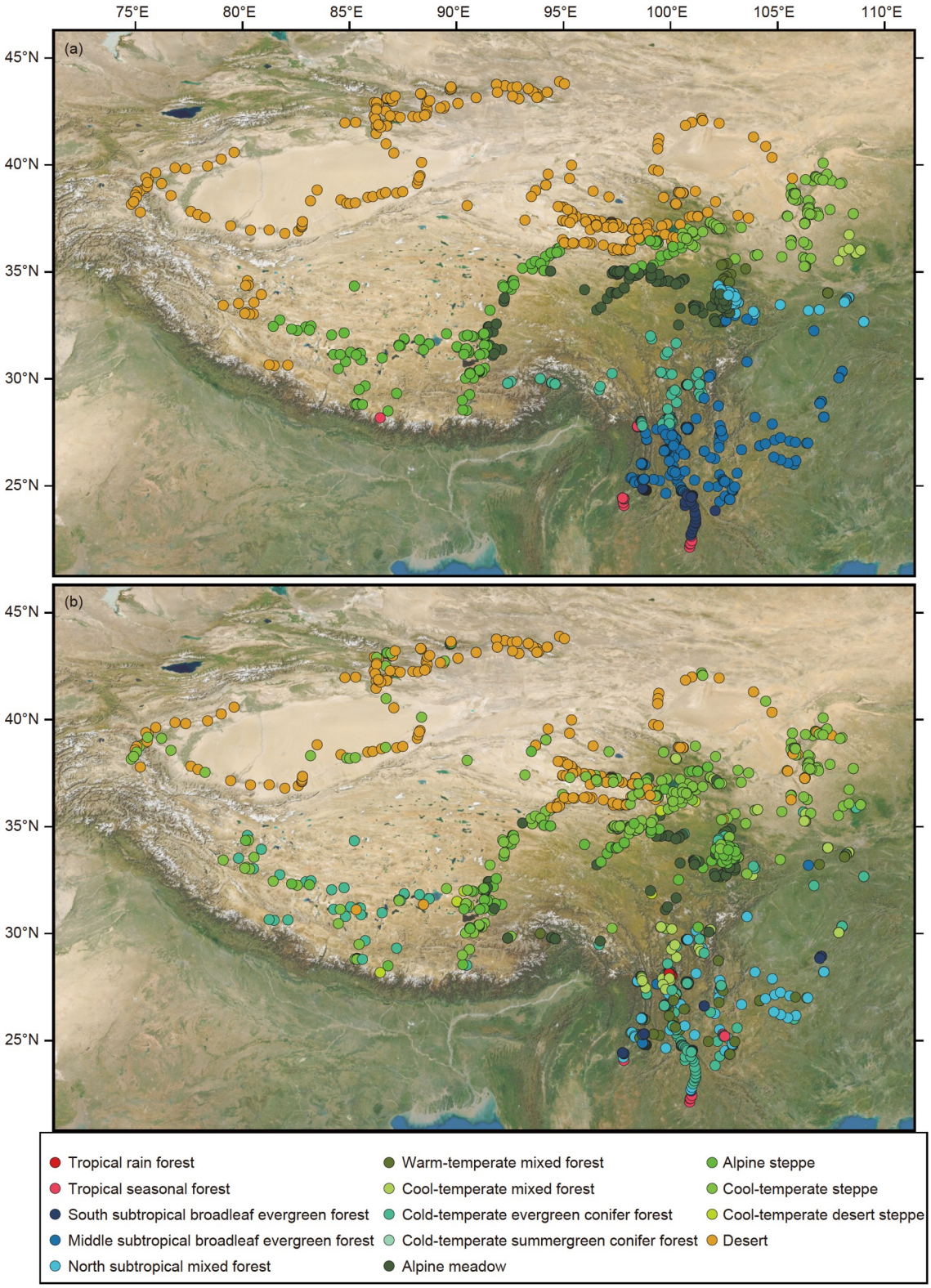

Figure 4 The observed modern biomes (a) and the predicted biomes produced via the biomisation method (b) for the modern pollen samples.

incorrectly assigned to the STFO and WAMF, respectively.

The other biome types showed unsatisfactory accuracies. Only 21.18% of the CLEC samples were correctly assigned, while the majority of the CLEC samples were assigned to COMX (26.47%), ALME (15.29%), and WAMF (12.94%). Nearly half of the ALST samples were incorrectly assigned to STEP (44.38%), and only 15.63% of the reconstructions were consistent with the observed vegetation. Large portions

Table 2 Confusion matrix of the biomisation method performances for classifying modern biomes using modern pollen assemblages[a)]

Observed vs. predicted	TRFO	TSFO	STFO	MTFO	WAMF	TEDE	COMX	CLEC	CLDC	ALME	STEP	TEDS	ALST	DESE	Accuracy
TRFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
TSFO	0	10	7	2	6	1	0	0	1	0	0	1	0	0	0.36
STFO	0	0	0	1	5	2	4	49	3	0	0	0	0	0	0.00
MTFO	6	13	15	13	222	45	48	292	1	6	4	2	1	0	0.02
WAMF	0	0	0	1	1	3	12	7	0	10	4	0	13	1	0.02
TEDE	0	0	0	0	0	2	7	3	0	4	6	0	1	1	0.08
COMX	0	0	0	0	0	0	1	1	0	0	6	0	0	0	0.13
CLEC	17	0	0	2	22	3	45	36	0	26	8	1	10	0	0.21
CLDC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
ALME	0	0	0	0	0	0	1	3	0	63	46	0	48	2	0.39
STEP	0	0	0	0	0	0	14	5	0	4	84	1	1	11	0.70
TEDS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
ALST	0	0	0	0	0	0	0	26	0	6	71	8	25	24	0.16
DESE	0	0	0	0	0	0	8	7	0	5	103	0	14	208	0.60

a) TRFO, tropical rain forest; TSFO, tropical seasonal forest; STFO, south subtropical broadleaf evergreen forest; MTFO, middle subtropical broadleaf evergreen forest; WAMF, north subtropical mixed forest; TEDE, warm-temperate mixed forest; COMX, cool-temperate mixed forest; CLEC, cold-temperate evergreen conifer forest; CLDC, cold-temperate summergreen conifer forest; ALME, alpine meadow; STEP, cool-temperate steppe; TEDS, cool-temperate desert steppe; ALST, alpine steppe; DESE, desert

of the ALST samples were incorrectly assigned to the CLEC (16.25%) and DESE (15.00%). Only one COMX sample was correctly reconstructed, and the others were largely misidentified as STEP (75%). The TEDE samples were identified as TEDE (8.33%), COMX (29.17%), STEP (25.00%), ALME (16.67%), and CLEC (12.50%). The MTFO biome had the largest number of samples but had a 2% reconstruction accuracy, with a large number of samples being misidentified as CLEC (43.71%) and WAMF (33.23%). Few of the WAMF samples were correctly reconstructed (1.92%), while the majority were misidentified as ALST (25.00%), COMX (23.08%), ALME (19.23%), and CLEC (13.46%). None of the STFO samples were correctly reconstructed, and most were misidentified as CLEC (76.56%).

3.3 Palaeovegetation reconstruction

The random forest model showed higher accuracy than the biomisation method based on comparing the reconstructions from modern pollen data with observed biomes. Therefore, the random forest model was adopted to reconstruct biome changes on the Tibetan Plateau from 22 ka BP to the present.

Past biome distributions of the Tibetan Plateau were demonstrated via the biome reconstructions of 51 fossil sites using the random forest model (serial maps were shown in Figure S1 in Appendix B). Five stages reflecting major changes in the biome distribution were determined, although fluctuations existed within each stage (Figure 5).

Stage I (22–17 ka BP). This period was characterised by widespread desert biomes across the Tibetan Plateau. Nearly all sites in the western, central, and northern parts of the Tibetan Plateau were reconstructed as temperate semi-shrubby and shrubby deserts or warm-temperate shrubby and semi-shrubby deserts in the 11 time-windows. In the eastern and southeastern areas, subtropical mountain cold-temperate needleleaf forests or middle subtropical broadleaf evergreen forests occurred. From 18.5–17 ka BP, several sites were identified as alpine steppes in the eastern area, and subalpine scrub & alpine meadow biomes occasionally occurred.

Stage II (16.5–12 ka BP). During this period, the temperate semi-shrubby and shrubby desert and warm-temperate shrubby and semi-shrubby desert biomes gradually retreated to the northern part of the Tibetan Plateau, and middle subtropical broadleaf evergreen forests remained in the eastern and southeastern areas. Alpine steppe became prevalent in the central and eastern regions, and an alpine steppe belt emerged. The subalpine scrub & alpine meadows occupied a small region in the southeastern Tibetan Plateau.

Stage III (11.5–8 ka BP). Temperate semi-shrubby and shrubby deserts and warm-temperate shrubby and semi-shrubby deserts mainly distributed in the northern and western portions of the Tibetan Plateau but also occurred occasionally at central and eastern Tibetan Plateau sites. Middle subtropical broadleaf evergreen forest mainly occurred in the eastern and southeastern areas, and occasionally occurred in southern and central Tibetan Plateau sites. A subalpine scrub and alpine meadow belt gradually formed in the eastern-southeastern area of the Tibetan Plateau, which consequently compressed the alpine steppe extents.

Stage IV (7.5–6 ka BP). During this period, temperate

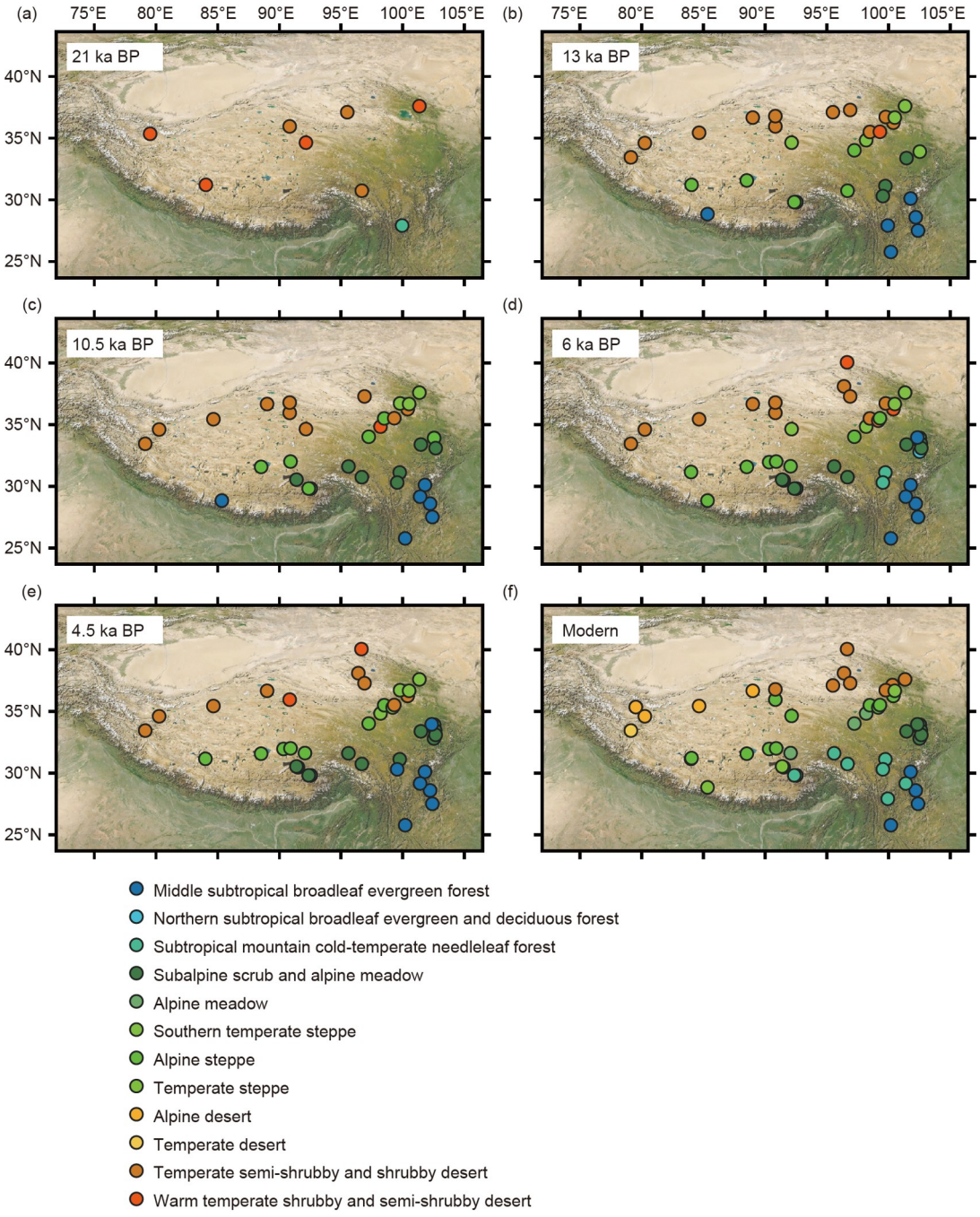

Figure 5 Selected maps representing the typical patterns of the five-stage changes in the Tibetan Plateau biome distributions over the last 22 ka reconstructed using the random forest model ((a)–(e), Stages I–V), and modern biomes (f) of the fossil sites.

semi-shrubby and shrubby deserts and warm-temperate shrubby and semi-shrubby deserts were generally restricted to the northern and western parts of the Tibetan Plateau. The subtropical mountain cold-temperate needleleaf and middle subtropical broadleaf evergreen forests expanded towards the southeastern part of the Tibetan Plateau at the expense of subalpine scrub & alpine meadow biome extents.

Stage V (5.5–0 ka BP). In this stage, temperate semi-shrubby and shrubby deserts and warm-temperate shrubby and semi-shrubby deserts remained in the northern and western areas; alpine steppes were dominant in the central part; and subalpine scrub & alpine meadow biomes often occurred in the eastern and southern regions of that the Tibetan Plateau. Subtropical mountain cold-temperate needleleaf and middle subtropical broadleaf evergreen forests generally occupied the southeastern region of Tibetan Plateau, and spread farther north in some time windows such as 4 and 1.5 ka BP.

4. Discussion

4.1 Robustness of the random forest model to reconstruct modern biomes

The random forest model showed a reliable performance in predicting modern biomes from modern pollen assemblages based on a comparison with the observed biomes (Table 1, Figure 2). However, the classification performance of the model differs for each biome.

Biomes in the meadow and desert categories all showed highly accurate classification abilities, including alpine meadow, subalpine scrub & alpine meadow, warm temperate shrubby and semi-shrubby desert, and temperate semi-shrubby and shrubby desert (Table 1). Within the forest category, the middle subtropical broadleaf evergreen forest and subtropical mountain cold-temperate needleleaf forest exhibited high prediction accuracies, and the southern temperate steppe was the only steppe category to show high accuracy.

Using the random forest model, the forest samples were easily misidentified as middle subtropical broadleaf evergreen forest (Table 1), which is primarily attributed to the similar pollen signals of arboreal taxa (e.g., *Alnus*, *Betula*, and *Quercus*) among the forest categories. Some of the forest samples were misidentified as non-forest biomes. For instance, many northern subtropical broadleaf evergreen and deciduous forest samples were assigned to the subalpine scrub & alpine meadow biome. Because the northern subtropical broadleaf evergreen and deciduous forest zone is located beside the subalpine scrub & alpine meadow zone (Figure 1), the samples located near the boundary are vulnerable to the pollen signals of neighbouring vegetation zones. For the same reason, the southern warm-temperate deciduous *Quercus* forest samples were often misidentified as the subalpine scrub & alpine meadow or southern temperate steppe.

The samples of alpine steppe and temperate steppe biomes were poorly identified by the random forest model (Table 1), with most of the misidentifications being assigned to southern temperate steppe, alpine meadow and temperate semi-shrubby and shrubby desert biomes. The inaccuracy was mainly attributed to the shared pollen components of these non-forest biomes. Additionally, a significant number of the alpine steppe and temperate steppe samples were incorrectly assigned to the middle subtropical broadleaf evergreen forest, which is resulted from long-distance transportation of arboreal pollen from subtropical forests in the southeastern and eastern regions.

Compared with the random forest model, the biomisation method had much lower accuracies for predicting modern biomes using the same modern pollen dataset (Table 2). Only cool-temperate steppe (STEP) and desert (DESE) showed reliable reconstructions. Samples of alpine steppe (ALST) and alpine meadow (ALME) were frequently misidentified as other non-forest biomes. The forest samples had even more ambiguous assignments. Tropical seasonal forest (TSFO) had relatively high accuracy, and misidentifications generally fall into other forest biomes. More than half of the north subtropical, warm-temperate, and cool-temperate mixed forest samples were identified as non-forest biomes. The cold-temperate evergreen conifer forest (CLEC) samples were assigned to 4–5 biomes. The south subtropical broadleaf evergreen forest (STFO) samples were all incorrectly assigned, and most of them were assigned to other forest biomes.

The unsatisfactory performance of the biomisation method may partly be attributed to the insufficient compatibility between biome classification scheme and vegetation zonation on the Tibetan Plateau. For instance, subalpine scrub and alpine meadows are important zonal vegetations in the eastern Tibetan Plateau, while they are merged into the alpine meadow in the biome classification scheme of Sun et al. (2020). Sun et al. (2020) reported higher reconstruction accuracy (80%) of the vegetation distributions across China based on modern pollen data using the biomisation method. This implies that with the available biome/PFT classification scheme, the biomisation method is suitable for subcontinental and global scale application; however, it is not suitable for biome distribution reconstruction of the Tibetan Plateau.

Reconstructions produced via the random forest model and biomisation method are not directly comparable because they utilise different biome classification approaches. The reconstructed biomes were grouped into mega-biomes according to bioclimate control similarities to enable a direct comparison between the two methods (Appendix B Table B4). Considering the mega-biomes, the overall accuracies of the random forest model (82.57%) and biomisation method (43.01%) were improved because the biomes sharing similar bioclimate controls and producing indistinguishable pollen signals were combined into one mega-biome at a coarser scale. Generally, the random forest model produced a more distinct zonal pattern in the study region than the biomisation method (Figure 6).

The random forest model produced clear spatial pattern for the mega-biomes of the modern pollen samples (Figure 6). Tropical forests occurred in the southernmost region of the study area, and subtropical forests distributed in the southeastern and eastern regions surrounding the Tibetan Plateau. Temperate forests occurred in a few locations in the eastern temperate region. Cool-temperate/boreal conifer forests distributed in the southeastern Tibetan Plateau. Alpine meadows/subalpine scrubs occupied the eastern area, and temperate/alpine deserts distributed across the northern region in and surrounding the Tibetan Plateau. Temperate/alpine steppes were scattered across the central Tibetan Plateau

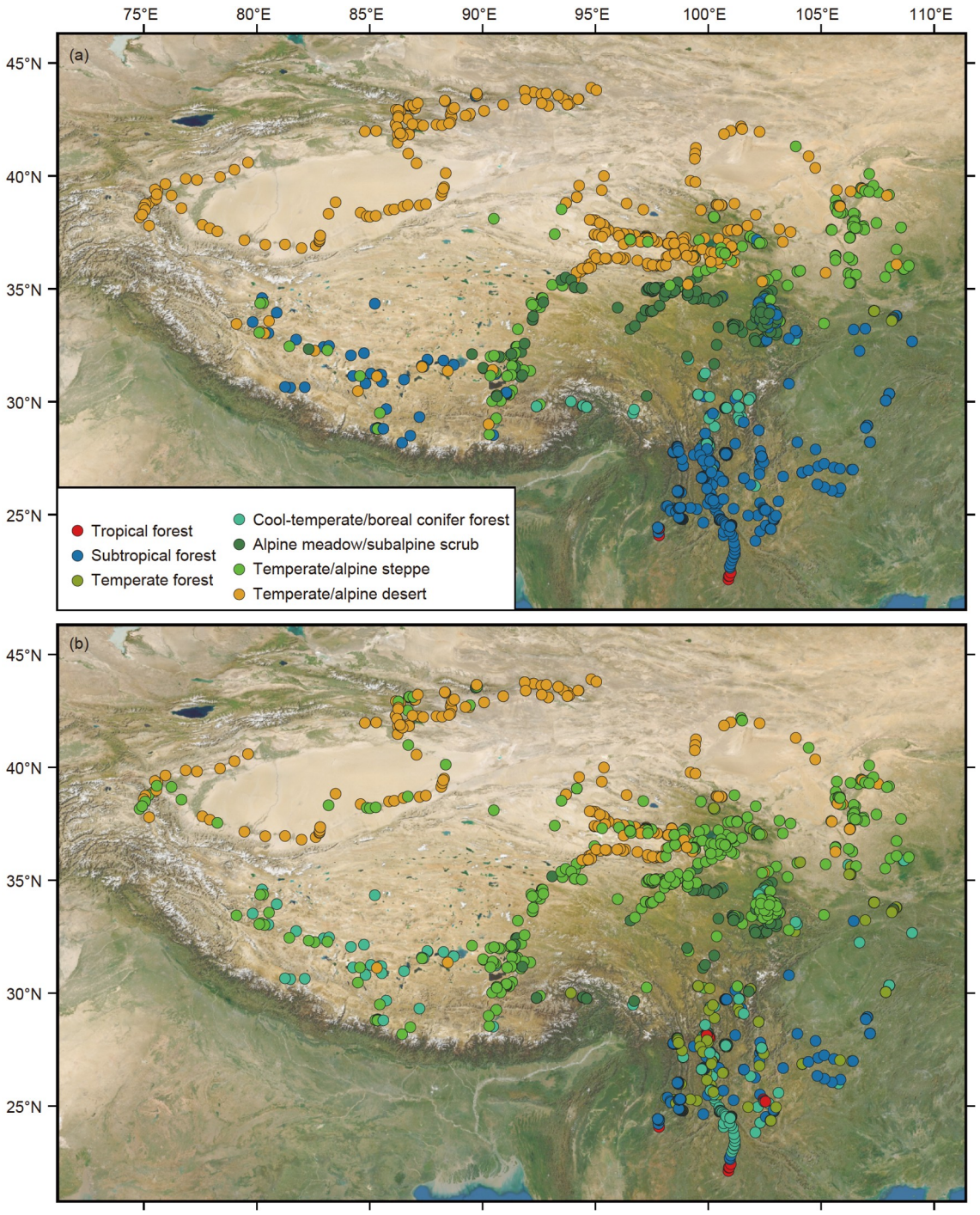

Figure 6 Comparison of modern mega-biomes reconstructed using the random forest model (a) and biomisation method (b) based on modern pollen data.

and the northeastern portion of the study region.

Results of biomisation method show a somewhat similar spatial pattern to that of the random forest model (Figure 6). Different forest mega-biomes occupied the areas surrounding the Tibetan Plateau to the east and southeast. Alpine meadows/subalpine scrubs distributed across the eastern Tibetan Plateau. Temperate/alpine steppes occupied a vast area across the entire Tibetan Plateau and the temperate region northeast of the plateau, while temperate/alpine deserts dominated the northern part of the study area.

Notably, many sites in the central and western part of the Tibetan Plateau were reconstructed as forest biomes by both methods, where alpine non-forest vegetation (alpine steppe and alpine desert) typically prevails. Long-distance pollen transportation, prominent in regions with minimal plant cover (Cour et al., 1999), from forests surrounding the Ti-

betan Plateau, is believed responsible for these incorrect reconstructions.

In summary, the random forest model shows great potential for reconstructing past biome changes on the Tibetan Plateau, while the biomisation method with a published biome/PFT classification scheme yielded unsatisfactory results.

4.2 Strengths and weaknesses of the two methods

The random forest algorithm is a newly introduced method for reconstructing biomes based on pollen data (Sobol and Finkelstein, 2018; Sobol et al., 2019). The biomisation method has frequently been used in previous palaeovegetation reconstruction studies (Prentice and Jolly, 2000; Pickett et al., 2004; Marchant et al., 2009) and is a reliable method for predicting modern and past biomes in China (Yu et al., 2000; Chen et al., 2010; Ni et al., 2010; Ni et al., 2014; Tian et al., 2018; Sun et al., 2020). Both methods have strengths and weaknesses, and each has specific error sources.

There are three advantages to applying the random forest algorithm to establish a pollen-biome reconstruction model (Sobol et al., 2019). First, pollen-biome relationships are established based on modern pollen data, thereby avoiding the subjective assignment of PFTs and biomes. Second, the entire pollen dataset is retained for analysis, which reduces information loss. Third, the resultant model is statistically validated and robust.

As a data driven approach, the robustness of the random forest model primarily relies on the quality of the training set. Pollen identification is one of the most important quality aspects of a pollen dataset. The pollen identification resolution can differ among the modern pollen samples, which are compilations of modern pollen data obtained from studies spanning decades. Higher taxonomic levels (genera or families) were generally used to homogenise the original pollen taxa of this study to maintain consistency, which increases the risk of involving different taxa characterised by different environmental adaptions into an individual pollen taxon.

The biases influencing the pollen-vegetation relationships also contribute to the uncertainty of the random forest model, such as the effect of long-distance transported pollen, differences of pollen productivities, variations among pollen assemblages within individual biome, and differences in pollen assemblages from different sediment types. In addition, the random forest algorithm is possible to identify ecologically irrelevant patterns in the dataset (Sobol et al., 2019).

Some weaknesses are attributed to the methodology. The random forest model cannot predict biome types outside of the biome range of the training set. For example, samples from the alpine desert and temperate desert zones of the Tibetan Plateau were excluded due to the limited number of sample sites (<15 sites); therefore, the random forest model cannot assign pollen assemblages to the two biomes. Moreover, a smaller sample size of a specific biome yields a weaker reconstruction performance for the biome. The seven biomes that exhibited high reconstruction accuracies constituted 81% of the modern pollen dataset (Table 1), and the remaining six biomes, which had low accuracies, constituted 19%. It's implied that small sample size may provide inadequate representation for the specific biome type.

The biomisation method has advantages of connecting pollen assemblage with biome based on ecological rules. PFT classification follows the ecophysiological and bioclimatic principles (Ni et al., 2010); the biomes are defined on a biogeographical basis; and the biome affinity scores are calculated using a rational equation (Prentice et al., 1996). Therefore, the robustness of the biomisation method primarily depends on the correlation between the biome/PFT classification scheme and the ecological foundations of the target vegetation with respect to palynology. Moreover, the biomisation method does not require a preliminary process to establish pollen-biome relationships based on the training set.

The biomisation method has three main weaknesses (Sobol et al., 2019). First, subjectivity is involved in defining PFTs and biomes and the selective removal of taxa from the dataset. Second, biomisation method relies on hand-tuning of model which may lead to overfitting. Third, a set of biome/PFT classification schemes is sensitive to vegetation change at limited ecological scale. For instance, the biome/PFT classification scheme of Sun et al. (2020) is robust for predicting biome distributions in China, but it exhibited a weak performance for the Tibetan Plateau biome reconstructions in this study. The adopted PFT/biome classification scheme was designed to fit the biome patterns across China, so it may not fully reflect the reginal-scale biome distributions of the Tibetan Plateau. It's expected that some adjustments in PFT and biome classifications based on regional consideration would improve the performance of biomisation method in future applications on the Tibetan Plateau.

The robustness of the random forest model could be improved by: (1) incorporating new modern pollen data from vegetation zones containing insufficient numbers of modern pollen assemblages (alpine desert zone and temperate desert zone); (2) improving the pollen data consistency and providing more appropriate modern analogues for fossil pollen assemblages from lake cores by performing additional modern pollen analyses on surface lake sediment; (3) reducing the influences of pollen representation and long-distance transported pollen by calibrating the pollen-vegetation relationship using the methods derived from the Extended R-value model (Bunting and Middleton, 2009; Sugita, 2007a,

2007b). The robustness of the biomisation method could be improved by designing a more compatible biome/PFT scheme for vegetation on the Tibetan Plateau.

4.3 Vegetation succession on the Tibetan Plateau since the LGM

This study illustrated the spatio-temporal changes of the Tibetan Plateau biome distributions from 22 ka BP to the present by using the pollen-based reconstruction of 51 sites via a random forest model (Figures 5 and Figure B1 in Appendix B). The fossil sites were scattered across the entire Tibetan Plateau and in different modern vegetation zones. The number of sites differs for the 45 time windows because the sediments of the sites span different time intervals, and no fossil sites could be found in some areas of the Tibetan Plateau in the specified time windows (see Figure S1 for details). The resultant biome distributions from the LGM to the present show that the biome pattern changes on the Tibetan Plateau since 22 ka BP generally correlated with global climate change and Asian monsoon dynamics.

Biomes on the Tibetan Plateau in Stage I (22–17 ka BP) were characterised by the constant dominance of deserts in most areas of the plateau, which corresponds to the cold and dry climate conditions during the LGM (Shakun and Carlson, 2010; Clark et al., 2012). Subtropical forests distributed in low-altitude regions along the southeastern and eastern margins. Between 18.5–17 ka BP, meadow and steppe biomes increased in the eastern region of the Tibetan Plateau, but they have not formed a zonal vegetation pattern to date.

In Stage II (16.5–12 ka BP), the steppe biomes expanded in the central and eastern Tibetan Plateau, and gradually developed a zonal steppe belt. The subtropical forest, steppe and desert biomes generally exhibited a latitudinal pattern from south to north. The meadow biome was more evident in the eastern portion of the Tibetan Plateau than in Stage I. Biome distributions of this stage are posited to be a response to climate change during the last deglaciation (Clark et al., 2012; Shi et al., 2021). A general amelioration of biomes on the Tibetan Plateau occurred from 16.5 to 13 ka BP with a gradual shrinking of the deserts and expansion of the steppes. At the end of this stage, the desert biome expanded south, probably corresponding to the Younger Dryas.

During the early Holocene (Stage III, 11.5–8 ka BP), the meadow biome began occupying the eastern and southeastern Tibetan Plateau, and gradually became zonal vegetation. The desert biome was restricted to the north and west, and the forest biome expanded toward the north. A forest-meadow-steppe-desert pattern was established during this stage from southeast to northwest, similar to that of the present Tibetan Plateau vegetation zonation. The biome changes were consistent with the early Holocene global climate-warming trend (Marcott et al., 2013) and Asian monsoon enhancement (Dykoski et al., 2005). Notably, that desert biome invaded the central Tibetan Plateau between 9.5 and 8 ka BP, which may be related to millennial-scale climate variations such as the 8.2 ka cooling event (Alley et al., 1997; Bond et al., 1997).

Biome distributions on the Tibetan Plateau during Stage IV (7.5–6 ka BP) revealed an "optimum" situation, which corresponds with the mid-Holocene climatic optimum of the East Asian monsoon region (Chen et al., 2019; Zhang et al., 2021). In this stage, subtropical forest biomes expanded to their northernmost extents observed in the study period, and the desert biomes were restricted to the northern and western regions.

In the last stage (Stage V, 5.5–0 ka BP), the subtropical forest biomes shifted south, and the meadows also expanded south. The steppe biomes occupied the modern alpine steppe zone and intermittently invaded the modern alpine meadow zone. The mid-to late Holocene climate generally showed a gradual deterioration (Chen et al., 2020; Marcott et al., 2013), and the Tibetan Plateau biomes experienced a corresponding change.

Noticeably, subtropical forest biomes were reconstructed at sites in the southern part of the Tibetan Plateau in multiple time windows, which are believed to be misidentifications by the random forest model attributed to two aspects: (1) the "pollution" of the long-distance transportation of arboreal pollen, particularly saccate pollen from the conifer forests in the southeastern Tibetan Plateau; and (2) relatively low prediction power of the random forest model for the alpine steppe and temperate steppe biomes in the central Tibetan Plateau (Table 1).

Generally, deserts covered most of the Tibetan Plateau during the LGM, and then gradually shifted north, occupying the northern and western regions of the Tibetan Plateau. Steppes started playing an important role at ~ 15 ka BP, and meadows expanded towards the eastern and southeastern portions of the Tibetan Plateau during the Holocene. Subtropical forests maintained dominance in eastern and southeastern regions throughout the study period, and expanded north during the early and mid-Holocene. Biomes in the eastern and southern parts of the Tibetan Plateau experienced the most frequent variations, which were exhibited as contests among the forest, meadow, and steppe biomes. Vegetation responded to climate change in different ways during different time intervals. For example, deserts expanded on the Tibetan Plateau during relatively cold and dry periods like LGM, Old Dryas and Younger Dryas, but they were restricted to the northern and western regions of the Tibetan Plateau after the mid-Holocene when the climate deteriorated.

The biome reconstructions using the random forest algorithm revealed the biome changes on the Tibetan Plateau

along a continuous timeline since the LGM at 500-year intervals. Comparisons between reconstructed biome changes and model simulations or pollen-based quantitative climate reconstructions in the future should provide new insights into the responses of alpine vegetation to climate changes. The strong correlation between the reconstructed biome changes on the Tibetan Plateau and contemporary global climate changes to some extent verified the reliability of random forest algorithms in reconstructing past biome changes. It's expected that the random forest algorithm should be a promising approach for quantitative biome reconstructions in other regions.

5. Conclusions

The lack of suitable methods hindered the quantitative biome reconstruction on the Tibetan Plateau since the LGM, which in turn hampered our understanding of alpine vegetation responses to climate changes and the model-proxy comparison to elucidate the responsible mechanisms for the vegetation changes. In this study, we introduced a supervised machine learning method to reconstruct past biome changes on the Tibetan Plateau. A pollen-based biome reconstruction model for the Tibetan Plateau was developed using the random forest algorithm based on the modern pollen assemblages. The random forest model had high accuracy for predicting modern biomes based on modern pollen data on the Tibetan Plateau and its vicinity. Comparison between random forest model and biomisation method indicated that the former exhibited a greater performance for predicting modern biomes in the study region, although both methods have strengths and weaknesses. Therefore, the random forest algorithm provides a valid tool to reconstruct past biome changes on the Tibetan Plateau using fossil pollen data.

The random forest model was used to reconstruct the Tibetan Plateau biome changes from 22 ka BP to the present based on fossil pollen records from 51 sites. A series of biome distribution maps of the Tibetan Plateau through time were constructed, and five stages reflecting the major changes of biome pattern over the last 22 ka were observed. The biome pattern changes on the Tibetan Plateau generally corresponded to global climate changes and Asian monsoon variations since the LGM. This study represents the first application of a machine learning method in palaeovegetation reconstruction in China. The good performance of the random forest algorithm in reconstructing past biome changes on the Tibetan Plateau implies that this method has great potential to be a powerful tool for quantitative biome reconstructions in other parts of China following the same procedure as this study. The quantitatively reconstructed biome changes on the Tibetan Plateau could provide evidence for model-proxy comparison, which would improve our understanding of the mechanism under alpine vegetation dynamics.

Acknowledgements *The authors wish to thank Prof. Zhuo ZHENG (Sun Yat-sen University, Guangzhou, China) and Prof. Xiayun XIAO (Nanjing Institute of Geography & Limnology, Chinese Academy of Sciences, Nanjing, China) for sharing modern pollen data, Dr Chen LIANG (Heibei GEO University, Shijiazhuang, China) for her assistance with data analysis, and Dr Zhiyong ZHANG (Lushan Botanical Garden, Chinese Academy of Sciences, Jiujiang, China) for the helpful discussions on biomisation method. This research was supported by the National Natural Science Foundation of China (Grant No. 41690113), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA20070101), the National Natural Science Foundation of China (Grant Nos. 42071114, 41977395, and 41671202).*

References

Alley R B, Mayewski P A, Sowers T, Stuiver M, Taylor K C, Clark P U. 1997. Holocene climatic instability: A prominent, widespread event 8200 yr ago. Geology, 25: 483–486

An Z S, Kutzbach J E, Prell W L, Porter S C. 2001. Evolution of Asian monsoons and phased uplift of the Himalaya-Tibetan plateau since Late Miocene times. Nature, 411: 62–66

An Z S, Wu G X, Li J P, Sun Y B, Liu Y, Zhou W J, Cai Y J, Duan A M, Li L, Mao J Y, Cheng H, Shi Z G, Tan L C, Yan H, Ao H, Chang H, Feng J. 2015. Global monsoon dynamics and climate change. Annu Rev Earth Planet Sci, 43: 29–77

Blaauw M, Christen J A. 2011. Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Anal, 6: 457–474

Blaauw M, Christen J A. 2019. rbacon: Age-Depth Modelling using Bayesian Statistics

Bond G, Showers W, Cheseby M, Lotti R, Almasi P, Demenocal P, Priore P, Cullen H, Hajdas I, Bonani G. 1997. A pervasive millennial-scale cycle in North Atlantic Holocene and Glacial climates. Science, 278: 1257–1266

Breiman L, Friedman J H, Olshen R A, Stone C J. 1984. Classification and Regression Trees. Monterey: Wadsworth and Brooks/Cole

Breiman L. 2001. Random Forests. Mach Learn, 45: 5–32

Bunting M J, Middleton R. 2009. Equifinality and uncertainty in the interpretation of pollen data: The Multiple Scenario Approach to reconstruction of past vegetation mosaics. Holocene, 19: 799–803

Cao X Y, Herzschuh U, Telford R J, Ni J. 2014. A modern pollen-climate dataset from China and Mongolia: Assessing its potential for climate reconstruction. Rev Palaeobot Palynol, 211: 87–96

Cao X Y, Ni J, Herzschuh U, Wang Y B, Zhao Y. 2013. A late Quaternary pollen dataset from eastern continental Asia for vegetation and climate reconstructions: Set up and evaluation. Rev Palaeobot Palynol, 194: 21–37

Chen F H, Chen J H, Huang W, Chen S Q, Huang X Z, Jin L Y, Jia J, Zhang X J, An C B, Zhang J W, Zhao Y, Yu Z C, Zhang R H, Liu J B, Zhou A F, Feng S. 2019. Westerlies Asia and monsoonal Asia: Spatiotemporal differences in climate change and possible mechanisms on decadal to sub-orbital timescales. Earth-Sci Rev, 192: 337–354

Chen F H, Zhang J F, Liu J B, Cao X Y, Hou J Z, Zhu L P, Xu X K, Liu X J, Wang M D, Wu D, Huang L X, Zeng T, Zhang S, Huang W, Zhang X, Yang K. 2020. Climate change, vegetation history, and landscape responses on the Tibetan Plateau during the Holocene: A comprehensive review. Quat Sci Rev, 243: 106444

Chen Y, Ni J, Herzschuh U. 2010. Quantifying modern biomes based on surface pollen data in China. Glob Planet Change, 74: 114–131

Clark P U, Shakun J D, Baker P A, Bartlein P J, Brewer S, Brook E, Carlson A E, Cheng H, Kaufman D S, Liu Z, Marchitto T M, Mix A C, Morrill C, Otto-Bliesner B L, Pahnke K, Russell J M, Whitlock C, Adkins J F, Blois J L, Clark J, Colman S M, Curry W B, Flower B P, He

F, Johnson T C, Lynch-Stieglitz J, Markgraf V, McManus J, Mitrovica J X, Moreno P I, Williams J W. 2012. Global climate evolution during the last deglaciation. Proc Natl Acad Sci USA, 109: E1134–E1142
Cour P, Zheng Z, Duzer D, Calleja M, Yao Z. 1999. Vegetational and climatic significance of modern pollen rain in northwestern Tibet. Rev Palaeobot Palynol, 104: 183–204
Cutler D R, Edwards Jr T C, Beard K H, Cutler A, Hess K T, Gibson J, Lawler J J. 2007. Random forests for classification in ecology. Ecology, 88: 2783–2792
Dallmeyer A, Claussen M, Herzschuh U, Fischer N. 2011. Holocene vegetation and biomass changes on the Tibetan Plateau—A model-pollen data comparison. Clim Past, 7: 881–901
Dykoski C A, Edwards R L, Cheng H, Yuan D X, Cai Y J, Zhang M L, Lin Y S, Qing J M, An Z S, Revenaugh J. 2005. A high-resolution, absolute-dated Holocene and deglacial Asian monsoon record from Dongge Cave, China. Earth Planet Sci Lett, 233: 71–86
Fall P L. 1992. Pollen accumulation in a montane region of Colorado, USA: A comparison of moss polsters, atmospheric traps, and natural basins. Rev Palaeobot Palynol, 72: 169–197
Felde V A, Peglar S M, Bjune A E, Grytnes J A, Birks H J B. 2014. The relationship between vegetation composition, vegetation zones and modern pollen assemblages in Setesdal, southern Norway. Holocene, 24: 985–1001
Guiot J, Goeury C. 1996. PPPBase, a software for statistical analysis of paleoecological and paleoclimatological data. Dendrochronol, 14: 295–300
Herzschuh U, Borkowski J, Schewe J, Mischke S, Tian F. 2014. Moisture-advection feedback supports strong early-to-mid Holocene monsoon climate on the eastern Tibetan Plateau as inferred from a pollen-based reconstruction. Palaeogeogr Palaeoclimatol Palaeoecol, 402: 44–54
Herzschuh U, Kramer A, Mischke S, Zhang C J. 2009. Quantitative climate and vegetation trends since the late glacial on the northeastern Tibetan Plateau deduced from Koucha Lake pollen spectra. Quat Res, 71: 162–171
Herzschuh U, Winter K, Wunnemann B, Li S J. 2006. A general cooling trend on the central Tibetan Plateau throughout the Holocene recorded by the Lake Zigetang pollen spectra. Quat Int, 154-155: 113–121
Li K, Liao M N, Ni J, Liu X Q, Wang Y B. 2019. Treeline composition and biodiversity change on the southeastern Tibetan Plateau during the past millennium, inferred from a high-resolution alpine pollen record. Quat Sci Rev, 206: 44–55
Liaw A, Wiener M. 2002. Classification and regression by randomForest. R News, 2: 18–22
Lisitsyna O V, Hicks S, Huusko A. 2012. Do moss samples, pollen traps and modern lake sediments all collect pollen in the same way? A comparison from the forest limit area of northernmost Europe. Veget Hist Archaeobot, 21: 187–199
Marchant R, Cleef A, Harrison S P, Hooghiemstra H, Markgraf V, van Boxel J, Ager T, Almeida L, Anderson R, Baied C, Behling H, Berrio J C, Burbridge R, Björck S, Byrne R, Bush M, Duivenvoorden J, Flenley J, De Oliveira P, van Geel B, Graf K, Gosling W D, Harbele S, van der Hammen T, Hansen B, Horn S, Kuhry P, Ledru M P, Mayle F, Leyden B, Lozano-García S, Melief A M, Moreno P, Moar N T, Prieto A, van Reenen G, Salgado-Labouriau M, Schäbitz F, Schreve-Brinkman E J, Wille M. 2009. Pollen-based biome reconstructions for Latin America at 0, 6000 and 18000 radiocarbon years ago. Clim Past, 5: 725–767
Marcott S A, Shakun J D, Clark P U, Mix A C. 2013. A reconstruction of regional and global temperature for the past 11,300 years. Science, 339: 1198–1201
Molnar P, Boos W R, Battisti D S. 2010. Orographic controls on climate and paleoclimate of Asia: Thermal and mechanical roles for the Tibetan Plateau. Annu Rev Earth Planet Sci, 38: 77–102
Ni J, Cao X Y, Jeltsch F, Herzschuh U. 2014. Biome distribution over the last 22,000 yr in China. Palaeogeogr Palaeoclimatol Palaeoecol, 409: 33–47
Ni J, Yu G, Harrison S P, Prentice I C. 2010. Palaeovegetation in China during the late Quaternary: Biome reconstructions based on a global scheme of plant functional types. Palaeogeogr Palaeoclimatol Palaeoecol, 289: 44–61
Pickett E J, Harrison S P, Hope G, Harle K, Dodson J R, Peter Kershaw A, Colin Prentice I, Backhouse J, Colhoun E A, D'Costa D, Flenley J, Grindrod J, Haberle S, Hassell C, Kenyon C, Macphail M, Martin H, Martin A H, McKenzie M, Newsome J C, Penny D, Powell J, Ian Raine J, Southern W, Stevenson J, Sutra J P, Thomas I, Kaars S, Ward J. 2004. Pollen-based reconstructions of biome distributions for Australia, Southeast Asia and the Pacific (SEAPAC region) at 0, 6000 and 18,000 ^{14}C yr BP. J Biogeogr, 31: 1381–1444
Prentice C, Guiot J, Huntley B, Jolly D, Cheddadi R. 1996. Reconstructing biomes from palaeoecological data: A general method and its application to European pollen data at 0 and 6 ka. Clim Dyn, 12: 185–194
Prentice I C, Jolly D. 2000. Mid-Holocene and glacial-maximum vegetation geography of the northern continents and Africa. J Biogeogr, 27: 507–519
Prentice I C, Webb III T. 1998. BIOME 6000: Reconstructing global mid-Holocene vegetation patterns from palaeoecological records. J Biogeogr, 25: 997–1005
Qin F. 2021. Modern pollen assemblages of the surface lake sediments from the steppe and desert zones of the Tibetan Plateau. Sci China Earth Sci, 64: 425–439
R Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
Shakun J D, Carlson A E. 2010. A global perspective on Last Glacial Maximum to Holocene climate change. Quat Sci Rev, 29: 1801–1816
Shen C M, Liu K B, Tang L Y, Overpeck J T. 2006. Quantitative relationships between modern pollen rain and climate in the Tibetan Plateau. Rev Palaeobot Palynol, 140: 61–77
Shi F, Lu H Y, Guo Z T, Yin Q Z, Wu H B, Xu C X, Zhang E L, Shi J F, Cheng J, Xiao X Y, Zhao C. 2021. The position of the Current Warm Period in the context of the past 22,000 years of summer climate in China. Geophys Res Lett, 48: e91940
Shi W, Jiang H C, Mao X, Xu H Y. 2020. Pollen record of climate change during the last deglaciation from the eastern Tibetan Plateau. PLoS ONE, 15: e0232803
Sobol M K, Finkelstein S A. 2018. Predictive pollen-based biome modeling using machine learning. PLoS ONE, 13: e0202214
Sobol M K, Scott L, Finkelstein S A. 2019. Reconstructing past biomes states using machine learning and modern pollen assemblages: A case study from Southern Africa. Quat Sci Rev, 212: 1–17
Song M H, Zhou C P, Ouyang H. 2005. Simulated distribution of vegetation types in response to climate change on the Tibetan Plateau. J Vegetation Sci, 16: 341–350
Sugita S. 2007a. Theory of quantitative reconstruction of vegetation I: Pollen from large sites REVEALS regional vegetation composition. Holocene, 17: 229–241
Sugita S. 2007b. Theory of quantitative reconstruction of vegetation II: All you need is LOVE. Holocene, 17: 243–257
Sun A Z, Luo Y L, Wu H B, Chen X D, Guo Z T. 2020. An updated biomization scheme and vegetation reconstruction based on a synthesis of modern and mid-Holocene pollen data in China. Glob Planet Change, 192: 103178
Tang L Y, Li C H. 2001. Temporal-spatial distribution of the Holocene vegetation in the Tibetan Plateau (in Chinese). J Glaciol Geocryol, 23: 367–374
Tang L, Shen C, Lu H, Li C, Ma Q. 2021. Fifty years of Quaternary palynology in the Tibetan Plateau. Sci China Earth Sci, 64: 1825–1843
Tian F, Cao X Y, Dallmeyer A, Lohmann G, Zhang X, Ni J, Andreev A, Anderson P M, Lozhkin A V, Bezrukova E, Rudaya N, Xu Q H, Herzschuh U. 2018. Biome changes and their inferred climatic drivers in northern and eastern continental Asia at selected times since 40 cal ka bp. Veget Hist Archaeobot, 27: 365–379
Wilmshurt J M, McGlone M S. 2005. Origin of pollen and spores in surface lake sediments: Comparison of modern palynomorph assemblages in moss cushions, surface soils and surface lake sediments. Rev Palaeobot

Palynol, 136: 1–15
Wu Z Y. 1980. Vegetation of China. Beijing: Science Press
Yu G, Chen X, Ni J, Cheddadi R, Guiot J, Han H, Harrison S P, Huang C, Ke M, Kong Z C, Li S, Li W Y, Liew P, Liu G, Liu J, Liu Q, Liu K B, Prentice I C, Qui W, Ren G, Song C, Sugita S, Sun X J, Tang L Y, van C E, Xia Y, Xu Q H, Yan S, Yang X, Zhao J, Zheng Z. 2000. Palaeovegetation of China: A pollen data-based synthesis for the mid-Holocene and last glacial maximum. J Biogeogr, 27: 635–664
Zhang X S. 1978. The plateau zonality of vegetation in Xizang (in Chinese). Acta Botanica Sin, 20: 140–149
Zhang X S. 2007. Vegetation Map of China and Its Geographic Pattern-Illustration of the Vegetation Map of The People's Republic of China (1:1000000) (in Chinese). Beijing: Geology Press
Zhang Y L, Li B Y, Zheng D. 2014. Datasets of the boundary and area of the Tibetan Plateau (in Chinese). Acta Geogr Sin, 69: 65–68
Zhang Z P, Liu J B, Chen J, Chen S Q, Shen Z W, Chen J, Liu X K, Wu D, Sheng Y W, Chen F H. 2021. Holocene climatic optimum in the East Asian monsoon region of China defined by climatic stability. Earth-Sci Rev, 212: 103450
Zhao Y, Liang C, Cui Q Y, Qin F, Zheng Z, Xiao X Y, Ma C M, Felde V A, Liu Y L, Li Q, Zhang Z Y, Herzschuh U, Xu Q H, Wei H C, Cai M T, Cao X Y, Guo Z T, Birks H J B. 2021. Temperature reconstructions for the last 1.74-Ma on the eastern Tibetan Plateau based on a novel pollen-based quantitative method. Glob Planet Change, 199: 103433
Zhao Y, Tzedakis P C, Li Q, Qin F, Cui Q, Liang C, Birks H J B, Liu Y L, Zhang Z Z, Ge J Y, Zhao H, Felde V A, Deng C L, Cai M T, Li H, Ren W H, Wei H C, Yang H F, Zhang J W, Yu Z C, Guo Z T. 2020. Evolution of vegetation and climate variability on the Tibetan Plateau over the past 1.74 million years. Sci Adv, 6: eaay6193
Zhao Y, Xu Q H, Huang X Z, Guo X L, Tao S C. 2009. Differences of modern pollen assemblages from lake sediments and surface soils in arid and semi-arid China and their significance for pollen-based quantitative climate reconstruction. Rev Palaeobot Palynol, 156: 519–524
Zhao Y, Yu Z C, Zhao W W. 2011. Holocene vegetation and climate histories in the eastern Tibetan Plateau: Controls by insolation-driven temperature or monsoon-derived precipitation changes? Quat Sci Rev, 30: 1173–1184
Zheng D, Zhang R Z, Yang Q. 1979. On the natural zonation in the Qinghai-Xizang Plateau. Acta Geogr Sin, 34: 1–11
Zheng Z, Wei J H, Huang K Y, Xu Q H, Lü H Y, Tarasov P, Luo C X, Beaudouin C, Deng Y, Pan A D, Zheng Y W, Luo Y L, Nakagawa T, Li C H, Yang S X, Peng H H, Cheddadi R. 2014. East Asian pollen database: Modern pollen distribution and its quantitative relationship with vegetation and climate. J Biogeogr, 41: 1819–1832

(Responsible editor: Haibin WU)