Received: 29 November 2022 | Revised: 17 January 2024 | Accepted: 9 March 2024

DOI: 10.1002/ldr.5109

RESEARCH ARTICLE

WILEY

Above-ground biomass retrieval with multi-source data: Prediction and applicability analysis in Eastern Mongolia

Shuxin Ji[1] | Batnyambuu Dashpurev[1] | Thanh Noi Phan[1] | Munkhtsetseg Dorj[2] | Yun Jäschke[3] | Lukas Lehnert[1]

[1]Department of Geography, Ludwig-Maximilian-University of Munich, Munich, Germany

[2]Agency of Land Administration and Management, Geodesy and Cartography, Government building XII, Barilgachidiin square, Ulaanbaatar, Mongolia

[3]Department of Botany, Senckenberg Museum of Natural History Görlitz, Görlitz, Germany

Correspondence
Shuxin Ji, Department of Geography, Ludwig-Maximilian-University of Munich, Luisenstr 37, 80333 Munich, Germany.
Email: shuxin.ji@iggf.geo.uni-muenchen.de

Funding information
Chinese Scholarship Council; German Federal Ministry of Education and Research, Grant/Award Number: 01LC1820B

Abstract

Grassland aboveground biomass (AGB) is a key variable to measure grassland productivity, and accurate assessment of AGB is important for optimizing grassland resource management and understanding carbon, water, and energy fluxes. Current approaches on large scales such as the Mongolian Steppe Ecosystem often combine field measurements with optical and/or synthetic aperture radar (SAR) data. Meanwhile, especially the representativeness of the field measurements for large-scale analysis have seldom been accounted for. Therefore, we provide the first remotely sensed AGB product for central and Eastern Mongolia which (1) uses random forest (RF), (2) is fully validated against over 600 field samples, and (3) applies a novel method, dissimilarity index (DI), to derive the area of applicability of the model with respect to the training data. Therefore, different remote sensing data sources such as multi-scale and multi-temporal optical images—Worldview 2 (WV2), Sentinel 2 (S2), and Landsat 8 (L8) in combination with SAR data are tested for their suitability to provide an area-wide estimation on large scale. The results showed that the AGB prediction by combining Sentinel 1 (S1) and S2 using RF had the highest accuracy. Furthermore, the model was applicable to at least 72.61% of the steppe area. Areas where the model was not applicable are mostly distributed along the edges of grassland. This study demonstrates the potential of combining Sentinel-derived indices and machine learning to provide a reliable AGB prediction for grassland for extremely large ecosystems with strong climatic gradients.

KEYWORDS
AGB, AOA, DI, machine learning, prediction, spectral and SAR

1 | INTRODUCTION

The Eurasian steppe, the largest continuous temperate grassland in the world, provides vital ecosystem services for humans. These include regulating global climate through significant organic carbon storage, sustaining biodiversity, and supporting livelihoods through agricultural and pastoral production (Bengtsson et al., 2019). However, the steppe is facing severe environmental pressure, climate change, and overexploitation of resources which contributes to the presumed grassland degradation and desertification, threatening the

integrity of this vast ecosystem (Darbalaeva et al., 2020; Dashpurev et al., 2020). Unsustainable use of grassland resources has also forced local communities to change land use, preventing improvements in their living standards in pastoral areas in the long run (Khishigbayar et al., 2015; Reid et al., 2014). The increase in the numbers of grazing, mining activities, as well as the effects of grassland reclamation for food production and development, have been the main causes of grassland degradation on the Mongolian Plateau (Leisher et al., 2012; Sainnemekh et al., 2022). In situ measurements showed that these human disturbances, coupled with climate warming, have reduced biodiversity and ecosystem functions within the region (Kauffman et al., 2021; Zhang et al., 2010). Therefore, timely and accurate monitoring of vegetation dynamics is key to assess potential grazing capacities and to protect the vast ecosystems (Wang et al., 2013). Since presumed drivers of vegetation conditions are changing over time, stable and validated time series products are required for Mongolia.

Aboveground biomass (AGB) is a key biophysical indicator characterizing grassland growth and conditions. Traditionally, data collected in situ within the enclosure in the undisturbed, natural plant communities is representative of aboveground net primary production. However, sampling with destructive methods is labor and cost-intensive and has limited spatial and temporal representativeness. Satellite remote sensing has been established as a low-cost and widely used tool to monitor vegetation across large areas and to provide accurate data for the management of vast ecosystems. Several previous studies reported the capability of optical data to be used for estimating AGB (Guerini Filho et al., 2020; Li et al., 2013; Mundava et al., 2014; Otgonbayar et al., 2019; Ren & Zhou, 2019), however, several limitations remain, including (1) spaceborne optical remote sensing is limited by clouds; (2) saturation of the relationship between AGB and surface reflectance at moderate to dense grassland vegetation; and (3) the spectral information is mainly from the top of canopy and ignores vertical vegetation structure. As a result, uncertainties of AGB estimates may vary in space and time according to the quality of the optical satellite data and the vegetation type.

One option to include information on the vertical structure of grassland vegetation is to use synthetic aperture radar (SAR data) which provides observations at a high spatial resolution in the order of tens of meters. In addition, the data are independent of clouds and solar illumination (Torres et al., 2012; Veloso et al., 2017). Microwaves are sensitive to the water content of vegetation and soil, consequently, SAR data have been proven to be a good supplement, especially for estimating the AGB of forests and crops (Blickensdörfer et al., 2022; Forkuor et al., 2020). In contrast, only a few studies used SAR data to estimate the AGB of grasslands (J. Wang, Xiao, et al., 2019), which is noticeable because a free data source of SAR data became available with Sentinel-1 (S1) since 2014.

To estimate ABG based on any satellite data, a transfer function is necessary. Most commonly, such transfer functions are established using statistical approaches which can be divided into parametric and non-parametric methods (Güneralp et al., 2014). Machine learning (ML) methods fall into the latter category. In a simplified perspective, ML can be seen as searching for parameter values through a large option space, guided by training data, to find a solution that optimizes a performance metric. Among the 26 articles published before the end of 2019 on the use of ML/deep learning to estimate AGB in grassland, random forest (RF), support vector machines (SVM), and artificial neural networks are the three most used (Morais et al., 2021). Popular advanced methods used to retrieve biomass via multivariable satellite data are RF (Jansen et al., 2019; Meyer et al., 2017; Ramoelo et al., 2015) and SVM (J. Wang, Xiao, et al., 2019; Wu et al., 2016) that have been evaluated as a valuable tool to be independent of multicollinearity among predictor variables.

In most of the existing studies, one ML method is used to train a model based on sample data, which is then applied to a larger study area without paying, attention to whether the prediction is equally valid across the study area also beyond the training points used for training. For instance, a recent study on salinity intrusion mapping in Vietnam's Mekong Delta, lacked training points at two cities in the north (Nguyen et al., 2021). Another example is a recent study on mapping rainfall in Eastern Asia, where no training data are included from outside of China (Zhang et al., 2021), but predictions are still made for these areas. In this case, how reliable is the accuracy of prediction in such areas without training points?

Thus, the objectives of this paper were to (1) develop a fully validated time series of AGB for Eastern Mongolia, (2) investigate the important contribution of different indices and bands to AGB mapping, and (3) evaluate in which area model predictions are reliable.

2 | MATERIALS AND METHODS

2.1 | Study area

The largest area of mostly intact steppe ecosystem in the world is found in Eastern Mongolia which is characterized by a close integration of social and natural processes at an altitude of 500–1300 m (Figure 1). Average temperatures range between around −4 and −8°C and vary strongly among years (Harris et al., 2020). Annual precipitation rarely exceeds 400 mm and is typically much lower in the south and central desert and steppe regions. The vegetation in Eastern Mongolia is mainly grouped as a steppe zone, and a limited area belongs to the forest-steppe belt (in the north and east), and a small area in the south is covered by desert steppe. Dry steppe is the most dominant steppe type in Eastern Mongolia (Tuvshintogtokh, 2014), while meadow steppe, mountain steppe, and desertified steppe occur in the area of the forest-steppe belt and transitional area to desert steppe. Our study region covers mainly dry steppe, and the most dominant grass communities are *Stipa krylovii* + *Leymus chinensis* + forbs and *Stipa grandis* + *Caragana* spp. + forbs. Most frequent common species include grass species such as *S. krylovii*, *S. grandis*, *L. chinensis*, and *Cleistogenes squarrosa*; sedge species like *Carex duriuscula*; shrub species like *Caragana microphylla*, *Caragana stenophylla*; subshrubs such as *Artemisia frigida*, *Artemisia adamsii*; and forb species including *Allium polyrhizum*, *Convolvulus ammannii*, *Chenopodium* spp., and *Astragalus* spp.

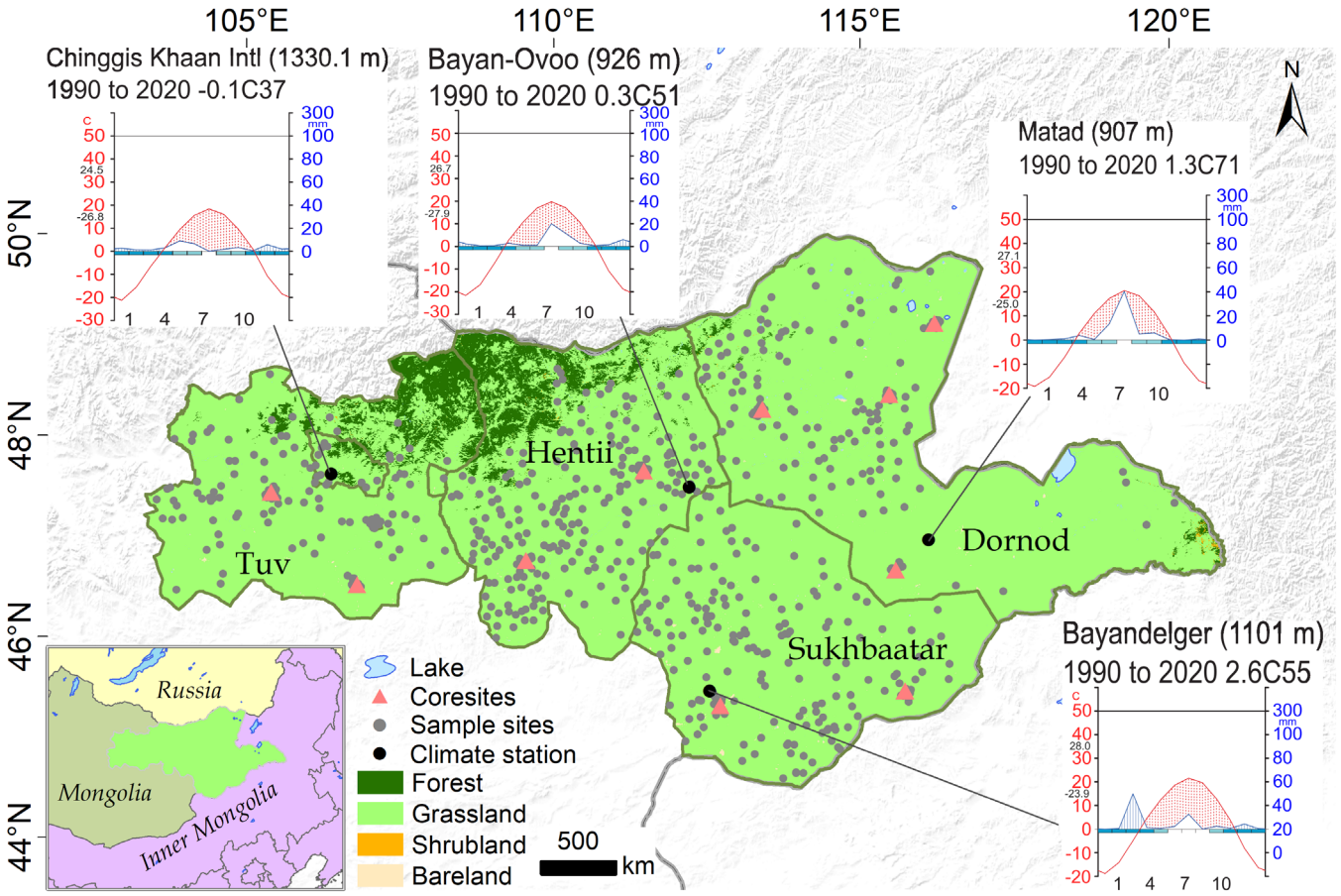

FIGURE 1 Overview of the location of the study area and land cover types (Phan et al., 2022). Climate diagrams for the period 1990–2020 are calculated using data from the Global Historical Climatology Network (Menne et al., 2012). [Colour figure can be viewed at wileyonlinelibrary.com]

2.2 | Data acquisition and pre-processing

In this section, we will first summarize the acquisition of field data used as a reference in the analysis. Then, we will describe the satellite data and the pre-processing used in this study. For a general overview of the methods applied, please refer to Figure 2. The major methodological steps include (1) data collection and preprocessing of different data sources (e.g., optical images from WV2, S2, and L8 and SAR data from S1); (2) creation of feature spaces (indices used as predictors derived from optical and radar sensor) to be used for AGB prediction; (3) comparing the performance of the two popular ML methods RF and SVM to retrieve the grassland biomass; and (4) evaluation of AGB estimation results. In this study, a total of 603 samples were systematically allocated on the grassland, of which 216 were from 10 core sites. Core sites differed from other sampling areas that WV2 data were available.

2.2.1 | In situ data collection

Field data on core sites were collected during the growing season of 2019 and 2021 by members from the MORE STEP project (https://www.morestep.org/), a stratified semi-random sampling method was conducted at different geographical scales to support the multiscale modeling and to ensure that only plots were considered which were representative for the surrounding vegetation. Semi-random in our context means that plots have been selected based on the constraint that surrounding vegetation must be homogeneous to reduce spatial scale effects. At 216 plots within the 10 core sites (Core sites differed from other sampling areas where WV2 data were available), AGB has been removed within a rectangle of 1 × 1 m marked with rulers (for statistics descriptive of field data see Table 1). Biomass samples have been dried and weighed in the laboratory. Field data outside of core sites were downloaded from the Agency for Land Administration and Management, Geodesy, and Cartography in Mongolia (https://egazar.gov.mn). The location of sampling plots was recorded with a GPS.

2.2.2 | Sentinel-1 data and pre-processing

S1 images were selected based on acquisition dates to minimize the time lapse between field campaigns and satellite overpasses and downloaded from the Copernicus Open Access Hub (https://scihub.copernicus.eu). The data were acquired in the Interferometric Wide Swath (IW) mode with dual polarization (VV, VH). Pre-processing includes four steps: radiometric calibration, speckle filtering, terrain

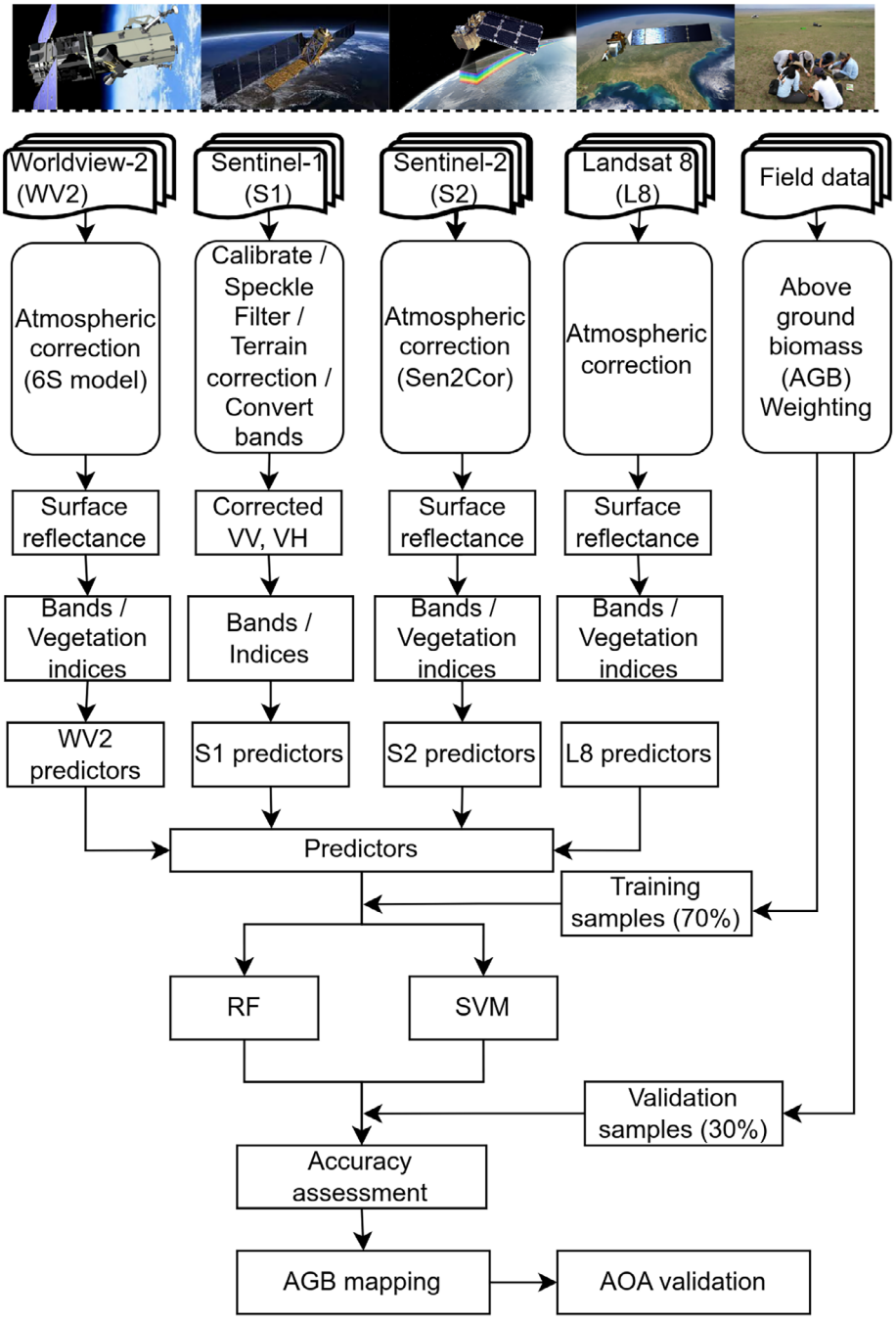

FIGURE 2 Schematic overview of AGB mapping with multi-source satellite data. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Descriptive statistics of the AGB (g/m^2) collected from core sites during the field campaign.

Year	N	Minimum	Mean	Median	Maximum	SD (%)	SE
2019	153	7.27	75.99	72.72	211.03	57.71	3.56
2021	63	17.85	96.18	94.17	200.16	39.95	4.88

correction, and conversion of the backscatter values to backscattering coefficients using the following equation.

$$\sigma_0(dB) = 10\log 10\sigma_0,$$

where $\sigma_0(dB)$ is the normalized radar cross section and σ_0 is the backscatter for a specific polarization, and the unit of the backscattering coefficient is dB. Shuttle Radar Topographic Mission Digital Elevation Model at 30 m resolution was used for the terrain correction. To reduce speckle noise, the Refined Lee speckle filtering algorithm (which adapts the window size to the local texture and edge information) was applied to the backscatter data which was selected due to its reported superior performance in SNAP (Lukin et al., 2018). To co-registrate of S1 and S2 datasets at a spatial resolution of 10 m, the S1 dataset was chosen as the reference layer and the bilinear interpolation method was utilized (De Luca, Silva, Di Fazio, & Modica, 2022).

2.2.3 | S2 data and pre-processing

S2 images acquired during June to August of 2019, 2020 and 2021 were downloaded from the ESA Sentinel Scientific Data Hub (http://scihub.copernicus.eu/) according to the closest dates for field samplings from June to August. The Level-1C was atmospherically corrected with Sen2Cor processor plugin. S2 has 13 spectral bands with different spatial resolutions from visible to short-wave infrared. Except bands 1, 9, and 10, all bands were pre-processed and included in the further analysis. The desired pixel size for the S2 image was selected to be 10 m. For the bands with a lower pixel resolution, the nearest neighbor resampling method was used.

2.2.4 | WV 2 data and pre-processing

WV2 images covered 10 core sites and were acquired in 2019 and 2021, scheduled in correspondence with the field campaigns. Radiance images were atmospherically corrected and transformed from top-of-atmosphere to bottom of the atmosphere reflectance via the 6S model (Vermote et al., 1997) adopted for large scales and altitudinal gradients (Curatola Fernández et al., 2015).

2.2.5 | L8 data and pre-processing

With extensive data archive and a wide range of wavelengths in the visible, near-infrared, and shortwave-infrared bands, Landsat images have been proven to be capable of predicting grass biomass with 30 m spatial resolution (Otgonbayar et al., 2019). L8 images were downloaded from the United States Geological Survey's Earth Explorer website (https://earthexplorer.usgs.gov/), and radiometric calibration and FLASSH atmospheric correction were conducted in ENVI 5.3. Only band 1 to band 7 were extracted in this study. For the summary of all satellite images used in this paper please see Table S1 in the supplementary material.

2.3 | Predictor selection and experiment design

For the variables from S1, in addition to the backscatter values, the difference (VV − VH), sum (VH + VV) (Vaglio Laurin et al., 2018), and ratio (VH/VV) (Veloso et al., 2017) were computed as predictors. Besides that, depending on the number of bands available in each optical sensor (S2, WV2, and L8), 12 common vegetation indices (CIre, mNDVIre, MSRre, MTCI, NDVIre, SRre, CIgreen, OSAVI, EVI2, mNDVI, MSR, and NDVI) (supplementary material Table S2) and all NDVI-like normalized differences indices (Thenkabail et al., 2000) (45 NDIs from S2, 32 from WV2, and 20 from L8) were calculated and used as additional—predictors. This resulted in feature spaces consisting of up to 72 predictor variables for S2 and S1 (for the number of predictors in other datasets see Table 2). In accordance with the objectives of the study, the performances of five different feature spaces were compared to understand the applicability of indices from different datasets and their combinations in mapping and predicting the AGB in Eastern Mongolia. R package "caret" (Kuhn, 2008) was employed to conduct two ML methods. Using the NDIs derived from two different spectral bands causes variable importance values to be difficult to interpret because one spectral band contributes to several predictors in the feature space. Consequently, the variable importance values were summed up by spectral bands contributing to each feature.

2.4 | Assessing the accuracy of model performance

To evaluate the accuracy of grass biomass prediction models in this study and reduce the error caused by the accidental division of training samples, 10-fold cross-validation was used. Each method in the model training was validated by a test dataset that was not used in the model training process in the same resolution (70% for training and 30% for testing), using the coefficient of determination (R^2), the cross-validation correlation coefficient (rcv), RMSE, relative RMSE (RMSEr), and bias for evaluating the accuracy of predicted values.

Then AOA was calculated to evaluate the representativeness of the model for areas not covered by the field sample locations (Meyer & Pebesma, 2021). Since the model has no knowledge about such areas, predictions on such areas must be considered more uncertain compared to areas covered by sufficient training data. We calculated the recently proposed dissimilarity index (DI) (based on the minimum distance to the training data in the multidimensional predictor space) which can be used to automatically derive the AOA of ML models. The basic idea of the calculation is based on the minimum distance between each pixel to the training data in the multidimensional predictor space, with predictors being weighted by their respective importance in the model. Then, AOA was derived by applying a threshold which was the maximum DI of the training data derived via cross-validation (Meyer & Pebesma, 2021).

3 | RESULTS

3.1 | Establishing model of AGB prediction

We used data from three satellite sensors to build models to predict AGB. In total, 6 AGB models (A, B, and C) have been developed based on variables from S1, S2 and their integration, and 4 models from WV2 (D) and L8 (E) with two ML methods (Table 3). The results showed that accuracies in predicting AGB differed among the models and ranged from moderate to high (r_{cv} between 0.46 and 0.75, RMSE between 935.83 and 532.12 kg/ha). Among the models, RF based on S1 and S2 (RF_S1 + S2) performed best with $r = 0.87$ and RMSE of 532.12 kg/ha. When compared only to optical sensor-based models, S2 showed better performance than WV2 and L8 (Figure 3). Irrespective of which ML methods has been used, adding S1 data as an additional predictor improved the accuracy. The comparison of the two

TABLE 2 Feature spaces in the ML regression models.

Predictors	Abbreviation	Description	Number
A: All optical data of S2	S2all	Spectral and indices	67
B: All SAR	S1all	VV, VH, and derivatives	5
C: Optical and SAR	S1S2all	All available predictors	72
D: All optical data WV2	WV2all	Spectral and indices	48
E: All optical data of L8	L8all	Spectral and indices	35

TABLE 3 Summary of biomass (kg/ha) prediction model results in 2019 and 2021.

Method	Predictors	RMSE (kg/ha)	r_{cv}	R^2	Bias	RMSEr
RF	A	535.12	0.86	0.74	14.16	40.56
	B	935.83	0.46	0.22	32.75	68.92
	C	532.12	0.87	0.75	16.75	40.79
	D	586.25	0.71	0.51	69.46	29.31
	E	610.49	0.80	0.65	18.89	66.93
SVM	A	582.71	0.82	0.67	39.21	44.41
	B	929.78	0.45	0.20	115.96	69.87
	C	604.03	0.82	0.67	33.73	46.04
	D	652.15	0.63	0.39	37.9	27.18
	E	638.53	0.79	0.62	62.11	76.78

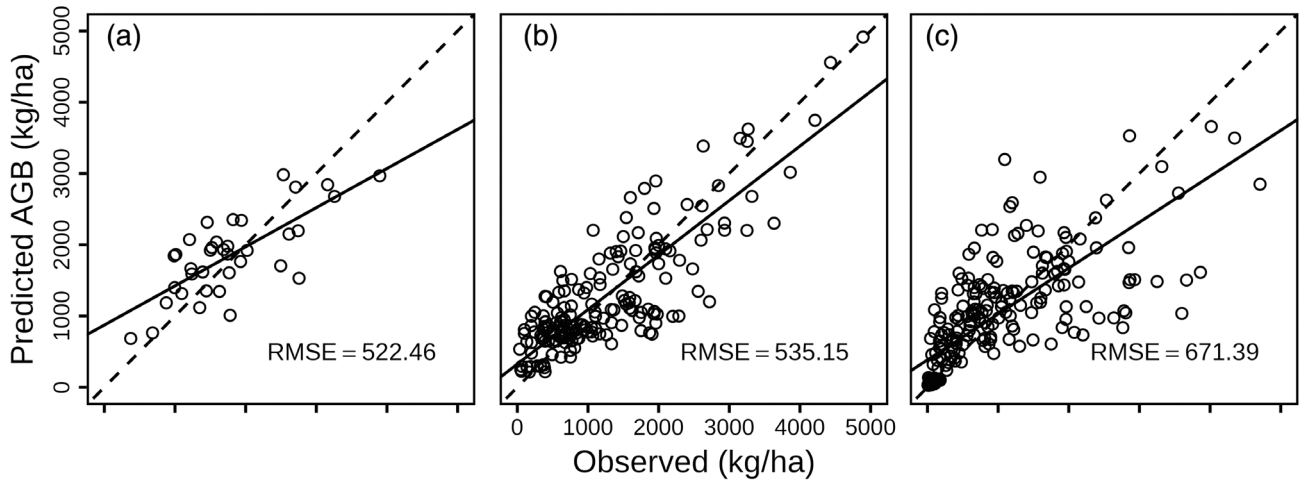

FIGURE 3 Scatterplot of predicted against observed AGB based on datasets of three optical sensors (see Table 2 for definitions of datasets), (a) WV2, (b) S2, and (c) L8. The dashed lines are the 1:1 line, and the solid lines are linear regressions. Note that there were only limited sample sites on the Worldview scale caused by the spatial extent of the available WV2 data.

ML methods showed that RF outperformed SVM in all cases except if only S1 data have been used. The performance of RF showed high quality when backscatter from S1 was introduced as additional predictors in the models. Models based on S1 and S2 ($r = 0.87$) still showed higher fits than models based on Landsat ($r = 0.80$).

3.2 | Variable importance

The sum of variable importance values per spectral band were highest in the red-edge part of the electromagnetic radiation for S2 and WV2 (Figure 4). For Landsat, the red band had the highest contribution to the model. The variable importance of the blue bands was lowest irrespective of the sensor used for AGB prediction. Intermediate importance values have been observed for near-infrared (NIR) and SWIR parts of the electromagnetic radiation.

3.3 | AGB prediction and validation

Figure 5 shows the average AGB estimation of Eastern Mongolia in June, July, and August (summer period), in 2021 as calculated by the best performing method (RF regression) on 10 m pixel resolution based on S1 and S2 data. In the north of Tuv, Khentii, and Dornod,

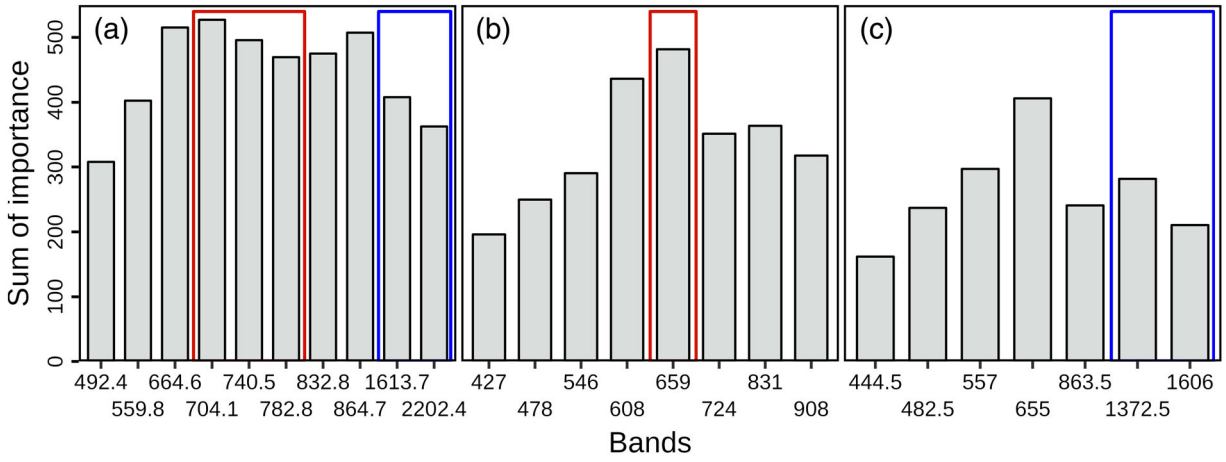

FIGURE 4 Sum of variable importance in three trained models, (a) S2, (b) WV2, (c) L8. Red rectangles show the red edge bands of S2 and WV2, and blue rectangles indicate the SWIR bands of S2 and L8. [Colour figure can be viewed at wileyonlinelibrary.com]

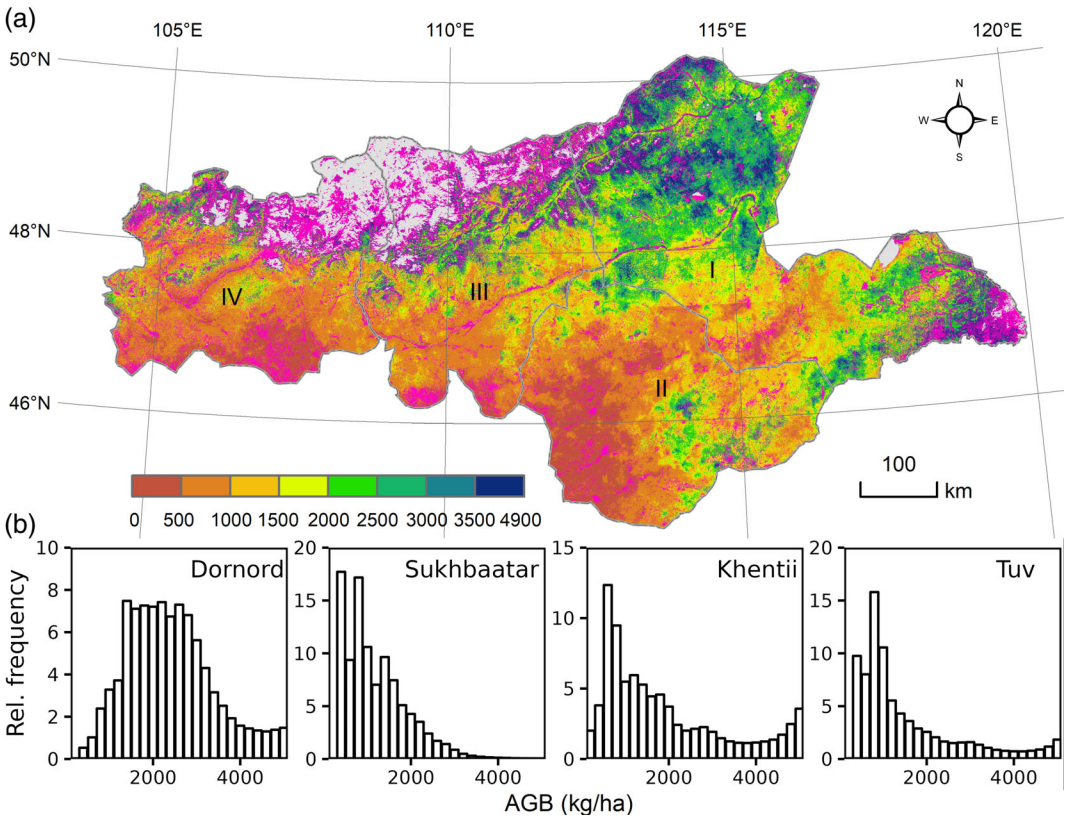

FIGURE 5 Application of the best-performing model to estimate biomass (in kilogram per hectare) across Eastern Mongolia (a) in 2021 (for 2019 and 2020, see Figure S1 in supplement). Areas not fulfilling AOA criteria are in pink, forest in gray have been excluded before prediction. Histograms in (b) show the distribution of AGB within different provinces (I, Dornod; II, Sukhbaatar; III, Khentii; IV, Tuv, locations are marked in a). [Colour figure can be viewed at wileyonlinelibrary.com]

AGB more than 4000 kg/ha were observed. This area is the transition zone between mountain steppes and forest steppes. Biomass is highest in the area of the most eastern part of Dornod. The histograms indicate the distribution of the AGB estimation in four provinces in the study area. The proportion of pixels with 1000–3000 kg/ha AGB in Dornod is more than 50%, mainly distributed in the eastern region. Around 40% of the grassland area in Sukhbaatar had less than 1000 kg/ha. In the north of Khentii and Tuv, models are not applicable to some areas because the distribution of meadow steppe even forest steppe, AGB below 500 kg/ha is predicted in the far south region, which is the area most closely to the Gobi desert in Mongolia.

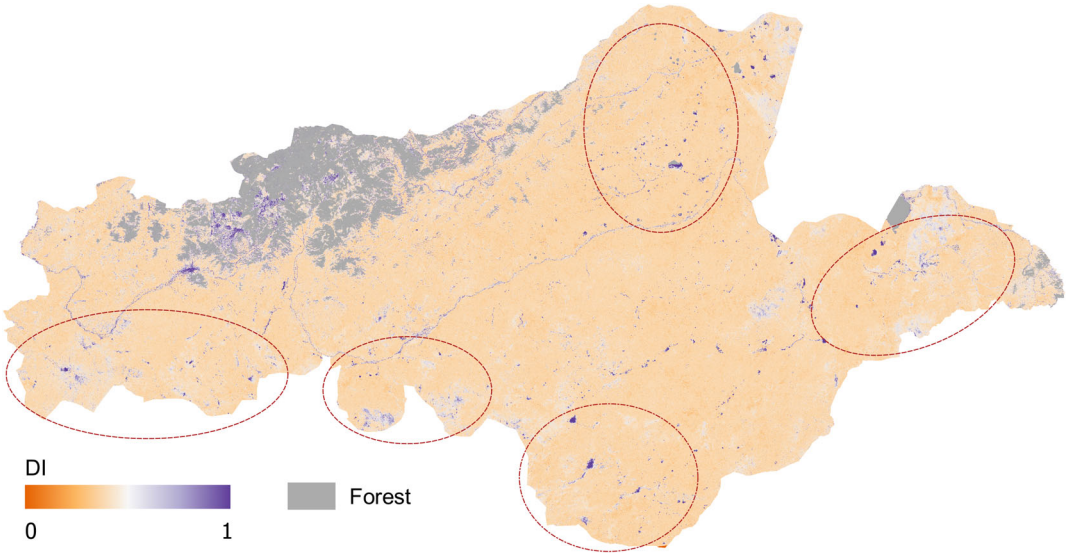

FIGURE 6 Dissimilarity index (DI) of the 2021 biomass prediction map, darker colors symbolize areas outside of AOA (For 2019 and 2020, see Figure S2 in supplement). Red ellipses mark regions with low AOAs discussed in the text. [Colour figure can be viewed at wileyonlinelibrary.com]

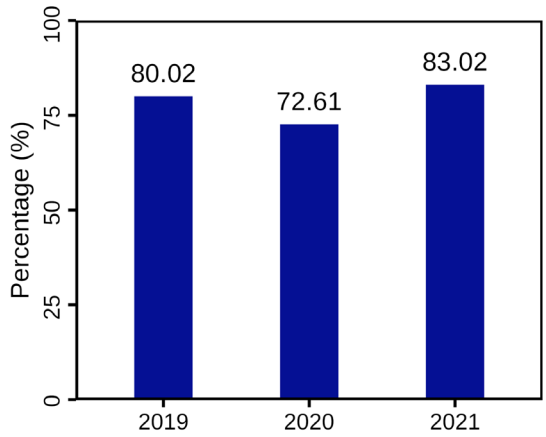

FIGURE 7 Percentage of pixels full-filling area of applicability (AOA) criteria for each year. [Colour figure can be viewed at wileyonlinelibrary.com]

Except for the area covered with forest steppe in the north part of Tuv and Khentii, the area outside of AOA mostly has AGB under 500 kg/ha or above 3500 kg/ha, especially the south part of the study area near the Gobi desert and north of Dornord where distributed mainly with mountain steppe (marked with red circles in Figure 6). Among the study periods, the spatial accuracy of the selected model remains above 72.61% (Figure 7), which means the model is trustworthy in the spatial prediction of grassland AGB in this region.

4 | DISCUSSION

Grassland AGB monitoring provides a valuable data source for management and decisions by local governments prohibiting degradation and, thus, maintaining biodiversity and ecosystem functioning. In this paper, we used different sources of satellite images for biomass estimation and found that the estimation accuracy under the combination of SAR and optical data with RF was promising.

4.1 | Accuracy of models

Integrating SAR data with optical data have been found to improve the accuracy of biomass estimates (De Luca, Silva, & Modica, 2022; De Luca, Silva, Di Fazio, & Modica, 2022; Malhi et al., 2022). However, our study found that integrating S1 and S2 only slightly improved the accuracy of grassland AGB prediction. This could be due to high percentages of bare soils in areas near desert steppe, leading to surface roughness and uncertainty of vegetation conditions, which can affect electromagnetic waves to some extent (Benninga et al., 2020). Additionally, the year 2019 was relatively wet in the study area, resulting in a high cloud cover. As a result, there were no cloud-free optical remote sensing images available near the time of field sampling, and only images acquired earlier in the year could be used. This resulted in a larger discrepancy between the vegetation condition at field samplings and the time of satellite image acquisition compared to other years.

Comparing different sensors, we found that integrating S1 and S2 data outperformed WV2 and L8. These sensors differ in spectral and spatial resolution. Using sums of variable importances per spectral band, we could show that red-edge bands contributed most to the models if such bands were available. In contrast, NIR bands were only intermediately important. SWIR bands were also intermediate if present. The red edge band is particularly useful for estimating chlorophyll content because it is sensitive to the absorption and reflectance properties of the pigment and reflectance is correlated with the

amount of chlorophyll in the vegetation, which is a key indicator of plant health and productivity (Tong & He, 2017). S2 has three red-edge bands (ranging from 689.1 to 802.8) which is different from WV2 (one band) and L8 (no bands in the red edge part). If SWIR bands are available, they contribute intermediately to the models. Here, the SWIR band at 1600 nm outperformed the bands at 2500 nm for S2 and L8 (WV2 does not have bands in the SWIR part of the electromagnetic radiation). From a physical perspective, bands in the SWIR region are particularly important for the discrimination of dead organic material from bare soils (Daughtry et al., 2005). Consequently, we conclude that S2 outperformed the other two sensors because of its spectral configuration encompassing three bands in the red edge and two bands in the SWIR part of the electromagnetic radiation. The higher spatial resolution of WV2 does not equalize the disadvantage of the lower spectral resolution of its sensor in homogeneous grasslands such as the Eastern Mongolian Steppe (Spagnuolo et al., 2020).

The integration of optical and radar sensors, along with the use of ML methods for biomass prediction, has become a popular practice due to its ability to produce more accurate results than traditional empirical models. However, spatial accuracy assessments are rarely performed, which is particularly important for large-scale analyses. In this study, we introduced a new approach to evaluate the model's ability to predict biomass at each pixel, and we found a minimal AOA of 72.61%, indicating that the training data used in this study is representative of most parts of the study area. Nevertheless, areas not represented by training data were mainly observed near the forest steppe in the north and the Gobi desert in the south, suggesting that future fieldwork should pay more attention to collecting samples from the ecological transition areas at the edges of the pristine steppe.

4.2 | Limitations and future work

AGB prediction through vegetation indices generated by optical sensors will cause "saturation", because these indices are constructed based on healthy vegetation absorbing radiation in a certain wavelength range and reflecting radiation in another, such as NDVI is based on the principle that healthy vegetation reflects more NIR radiation and absorbs more red radiation than non-vegetation surfaces, such as soil or water. Limited by the spectral radiation range of sensors, "saturation" is difficult to completely eliminate. In the southwest of Eastern Mongolia, where desert grassland dominates and the surface is more exposed, the prediction may be ignored or decreased, which have been proved optical indices to be sensitive to soil optical properties under conditions of incomplete vegetation cover (Ren et al., 2018; G. Wang, Liu, et al., 2019). While in the eastern Dornod province, where meadow steppe dominates, relying solely on the indices may lead to a lower estimated biomass. Previous studies have utilized hyperspectral sensors, which can better capture the canopy reflection signal and mitigate the influence of bare soil (Cooper et al., 2021; Zandler et al., 2015). Therefore, our direction is to develop a new ML model for biomass inversion and correction, using data from the upcoming CHIME hyperspectral satellite data.

5 | CONCLUSION

The objectives of this study were threefold: First, we aimed to develop the first fully validated time series of AGB for Eastern Mongolian Steppes. Therefore, we used over 600 in situ samples to train powerful ML models. In addition, we compared the suitability of four different sensors to estimate the AGB of Mongolian grasslands and found that combining S1 and S2 outperformed models solely based on L8 data or the high spatial resolution data of WV2. Second, the importance of different spectral regions for AGB mapping was evaluated. Here, we found the red edge band is particularly useful for estimating chlorophyll content, the high spatial resolution of WV2 seems not advantageous compared to the additional spectral bands of S2 especially in the red edge and SWIR. Third, we aimed to evaluate, which spatial areas model predictions are reliable based on the configuration of in situ samples. We found that the prediction accuracy of the model is higher when the sampling point is in a more homogeneous grassland. This means that uncertainties of AGB estimates are low in the north near to forest steppe and south covered by the Gobi desert, while they are increasing in the typical steppe.

Sampling data acquisition often has accessibility limitations, resulting in a limited number of ground sampling points. Therefore, it is crucial to ensure the accuracy of these points to facilitate ML-based regional predicting. To achieve this, accurate decision-making using AOA and DI is essential. As we strive for continuity in future sampling work, it is important not to overlook sample collection in diverse environments.

AUTHOR CONTRIBUTIONS

Shuxin Ji carried out the research and wrote the first version of the manuscript. Batnyambuu Dashpurev did experimental support and data acquisition. Munkhtsetseg Dorj helped collect and process the field data. Thanh Noi Phan conducted the processing of the big scenes dataset. Yun Jäschke provides vegetation type and species information. Lukas Lehnert provided the scientific guidance and code contribution.

ACKNOWLEDGMENTS

This study was funded by German Federal Ministry of Education and Research (01LC1820B), German Academic Exchange Service, and Chinese Scholarship Council. Thanks go to the MORE STEP project for supporting this study. We are grateful that the Sentinel-1 and Sentinel-2 data could be downloaded without charge from European Space Agency and Landsat 8 data can be downloaded from USGS. Special thanks go to Agency for Land Management in Mongolia for sharing field sample data. We are grateful for the open-source application SNAP developed by European Spatial Agency and two anonymous reviewers for their comments that improved our paper. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Shuxin Ji https://orcid.org/0000-0003-4081-5506

Thanh Noi Phan https://orcid.org/0000-0002-2747-5028

REFERENCES

Benninga, H.-J. F., van der Velde, R., & Su, Z. (2020). Sentinel-1 soil moisture content and its uncertainty over sparsely vegetated fields. *Journal of Hydrology X*, *9*, 100066. https://doi.org/10.1016/j.hydroa.2020.100066

Bengtsson, J., Bullock, J. M., Egoh, B., Everson, C., Everson, T., O'Connor, T., O'Farrell, P. J., Smith, H. G., & Lindborg, R. (2019). Grasslands—More important for ecosystem services than you might think. *Ecosphere*, *10*, e02582.

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., & Hostert, P. (2022). Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. *Remote Sensing of Environment*, *269*, 112831. https://doi.org/10.1016/j.rse.2021.112831

Cooper, S., Okujeni, A., Pflugmacher, D., van der Linden, S., & Hostert, P. (2021). Combining simulated hyperspectral EnMAP and Landsat time series for forest aboveground biomass mapping. *International Journal of Applied Earth Observation and Geoinformation*, *98*, 102307. https://doi.org/10.1016/j.jag.2021.102307

Curatola Fernández, G., Obermeier, W., Gerique, A., Fernanda, M., López Sandoval, M., Lehnert, L., Thies, B., & Bendix, J. (2015). Land cover change in the Andes of southern Ecuador—Patterns and drivers. *Remote Sensing*, *7*, 2509–2542. https://doi.org/10.3390/rs70302509

Darbalaeva, D., Mikheeva, A., & Zhamyanova, Y. (2020). The socio-economic consequences of the desertification processes in Mongolia. *E3S Web of Conferences*, *164*, 11001. https://doi.org/10.1051/e3sconf/202016411001

Dashpurev, B., Bendix, J., & Lehnert, L. (2020). Monitoring oil exploitation infrastructure and dirt roads with object-based image analysis and random forest in the Eastern Mongolian Steppe. *Remote Sensing, EI*, *12*-(January), 144. https://doi.org/10.3390/rs12010144

Daughtry, C. S. T., Hunt, E. R., Doraiswamy, P. C., & McMurtrey, J. E. (2005). Remote sensing the spatial distribution of crop residues. *Agronomy Journal*, *97*, 864–871. https://doi.org/10.2134/agronj2003.0291

De Luca, G., Silva, J. M. N., & Modica, G. (2022). Regional-scale burned area mapping in Mediterranean regions based on the multitemporal composite integration of Sentinel-1 and Sentinel-2 data. *GIScience Remote Sensing*, *59*, 1678–1705. https://doi.org/10.1080/15481603.2022.2128251

De Luca, G., Silva, M. N., Di Fazio, S., & Modica, G. (2022). Integrated use of Sentinel-1 and Sentinel-2 data and open-source machine learning algorithms for land cover mapping in a Mediterranean region. *European Journal of Remote Sensing*, *55*, 52–70. https://doi.org/10.1080/22797254.2021.2018667

Forkuor, G., Benewinde Zoungrana, J.-B., Dimobe, K., Ouattara, B., Vadrevu, K. P., & Tondoh, J. E. (2020). Above-ground biomass mapping in west African dryland forest using Sentinel-1 and 2 datasets - a case study. *Remote Sensing Environment*, *236*, 111496. https://doi.org/10.1016/j.rse.2019.111496

Guerini Filho, M., Kuplich, T. M., & Quadros, F. L. F. D. (2020). Estimating natural grassland biomass by vegetation indices using sentinel 2 remote sensing data. *International Journal of Remote Sensing*, *41*, 2861–2876. https://doi.org/10.1080/01431161.2019.1697004

Güneralp, I., Filippi, A. M., & Randall, J. (2014). Estimation of floodplain aboveground biomass using multispectral remote sensing and nonparametric modeling. *International Journal of Applied Earth Observation and Geoinformation*, *33*, 119–126. https://doi.org/10.1016/j.jag.2014.05.004

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*, 109. https://doi.org/10.1038/s41597-020-0453-3

Jansen, V., Kolden, C., Greaves, H., & Eitel, J. (2019). Lidar provides novel insights into the effect of pixel size and grazing intensity on measures of spatial heterogeneity in a native bunchgrass ecosystem. *Remote Sensing Environment*, *235*, 111432. https://doi.org/10.1016/j.rse.2019.111432

Kauffman, M. J., Cagnacci, F., Chamaillé-Jammes, S., Hebblewhite, M., Hopcraft, J. G. C., Merkle, J. A., Mueller, T., Mysterud, A., Peters, W., Roettger, C., Steingisser, A., Meacham, J. E., Abera, K., Adamczewski, J., Aikens, E. O., Bartlam-Brooks, H., Bennitt, E., Berger, J., Boyd, C., … Zuther, S. (2021). Mapping out a future for ungulate migrations. *Science*, *372*, 566–569. https://doi.org/10.1126/science.abf0998

Khishigbayar, J., Fernández-Giménez, M. E., Angerer, J. P., Reid, R. S., Chantsallkham, J., Baasandorj, Y., & Zumberelmaa, D. (2015). Mongolian rangelands at a tipping point? Biomass and cover are stable but composition shifts and richness declines after 20 years of grazing and increasing temperatures. *Journal of Arid Environments*, *115*, 100–112. https://doi.org/10.1016/j.jaridenv.2015.01.007

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Leisher, C., Hess, S., Boucher, T. M., Beukering, P., & van Sanjayan, M. (2012). Measuring the impacts of community-based grasslands Management in Mongolia's Gobi. *PLoS One*, *7*, e30991. https://doi.org/10.1371/journal.pone.0030991

Li, F., Jiang, L., Wang, X., Zhang, X., Zheng, J., & Zhao, Q. (2013). Estimating grassland aboveground biomass using multitemporal MODIS data in the west Songnen plain, China. *Journal of Applied Remote Sensing*, *7*, 073546. https://doi.org/10.1117/1.JRS.7.073546

Lukin, V., Rubel, O., Kozhemiakin, R., Abramov, S., Shelestov, A., Lavreniuk, M., Meretsky, M., Vozel, B., Chehdi, K., 2018. *Despeckling of Multitemporal Sentinel SAR Images and its Impact on Agricultural Area Classification. Recent Advances and Applications in Remote Sensing*. https://doi.org/10.5772/intechopen.72577

Malhi, R. K. M., Anand, A., Srivastava, P. K., Chaudhary, S. K., Pandey, M. K., Behera, M. D., Kumar, A., Singh, P., & Sandhya Kiran, G. (2022). Synergistic evaluation of sentinel 1 and 2 for biomass estimation in a tropical forest of India. *Advances in Space Research*, *69*, 1752–1767. https://doi.org/10.1016/j.asr.2021.03.035

Menne, M., Durre, I., Vose, R., Gleason, B., & Houston, T. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*, 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Meyer, H., Lehnert, L. W., Wang, Y., Reudenbach, C., Nauss, T., & Bendix, J. (2017). From local spectral measurements to maps of vegetation cover and biomass on the Qinghai-Tibet-plateau: Do we need hyperspectral information? *International Journal of Applied Earth Observation and Geoinformation*, *55*, 21–31. https://doi.org/10.1016/j.jag.2016.10.001

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*, 1620–1633. https://doi.org/10.1111/2041-210X.13650

Morais, T. G., Teixeira, R. F. M., Figueiredo, M., & Domingos, T. (2021). The use of machine learning methods to estimate aboveground biomass of grasslands: A review. *Ecological Indicators*, *130*, 108081. https://doi.org/10.1016/j.ecolind.2021.108081

Mundava, C., Helmholz, P., Schut, A. G. T., Corner, R., McAtee, B., & Lamb, D. W. (2014). Evaluation of vegetation indices for rangeland biomass estimation in the Kimberley area of Western Australia. *ISPRS Annals of the Photogrammetry, Remote Sensing, II7*, 47–53. https://doi.org/10.5194/isprsannals-II-7-47-2014

Nguyen, T. G., Tran, N. A., Vu, P. L., Nguyen, Q.-H., Nguyen, H. D., & Bui, Q.-T. (2021). Salinity intrusion prediction using remote sensing and machine learning in data-limited regions: A case study in Vietnam's Mekong Delta. *Geoderma Regional*, *27*, e00424. https://doi.org/10.1016/j.geodrs.2021.e00424

Otgonbayar, M., Atzberger, C., Chambers, J., & Damdinsuren, A. (2019). Mapping pasture biomass in Mongolia using partial least squares, random Forest regression and Landsat 8 imagery. *International Journal of Remote Sensing*, *40*, 3204–3226. https://doi.org/10.1080/01431161.2018.1541110

Phan, T.-N., Dashpurev, B., Wiemer, F., & Lehnert, L. W. (2022). A simple, fast, and accurate method for land cover mapping in Mongolia. *Geocarto International*, *37*, 1–19. https://doi.org/10.1080/10106049.2022.2087759

Ramoelo, A., Cho, M. A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z., & Wolff, E. (2015). Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data. *International Journal of Applied Earth Observation and Geoinformation*, Special Issue on "Advances in remote sensing of vegetation function and traits", *43*, 43–54. https://doi.org/10.1016/j.jag.2014.12.010

Reid, R. S., Fernández-Giménez, M. E., & Galvin, K. A. (2014). Dynamics and resilience of rangelands and pastoral peoples around the globe. *Annual Review of Environment and Resources, EI*, *39*(1), 217–242. https://doi.org/10.1146/annurev-environ-020713-163329

Ren, H., & Zhou, G. (2019). Estimating green biomass ratio with remote sensing in arid grasslands. *Ecological Indicators*, *98*, 568–574. https://doi.org/10.1016/j.ecolind.2018.11.043

Ren, H., Zhou, G., & Zhang, F. (2018). Using negative soil adjustment factor in soil-adjusted vegetation index (SAVI) for aboveground living biomass estimation in arid grasslands. *Remote Sensing of Environment*, *209*, 439–445. https://doi.org/10.1016/j.rse.2018.02.068

Sainnemekh, S., Barrio, I. C., Densambuu, B., Bestelmeyer, B., & Aradóttir, Á. L. (2022). Rangeland degradation in Mongolia: A systematic review of the evidence. *Journal of Arid Environments*, *196*, 104654. https://doi.org/10.1016/j.jaridenv.2021.104654

Spagnuolo, O. S. B., Jarvey, J. C., Battaglia, M. J., Laubach, Z. M., Miller, M. E., Holekamp, K. E., & Bourgeau-Chavez, L. L. (2020). Mapping Kenyan Grassland Heights across large spatial scales with combined optical and radar satellite imagery. *Remote Sensing*, *12*, 1086. https://doi.org/10.3390/rs12071086

Thenkabail, P. S., Smith, R. B., & De Pauw, E. (2000). Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment*, *71*, 158–182. https://doi.org/10.1016/S0034-4257(99)00067-X

Tong, A., & He, Y. (2017). Estimating and mapping chlorophyll content for a heterogeneous grassland: Comparing prediction power of a suite of vegetation indices across scales between years. *ISPRS Journal of Photogrammetry and Remote Sensing*, *126*, 146–167. https://doi.org/10.1016/j.isprsjprs.2017.02.010

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., … Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9–24. https://doi.org/10.1016/j.rse.2011.05.028

Tuvshintogtokh, I., (2014). *The steppe vegetation of Mongolia*. Bembi san.

Vaglio Laurin, G., Balling, J., Corona, P., Mattioli, W., Papale, D., Puletti, N., Rizzo, M., Truckenbrodt, J., & Urban, M. (2018). Above-ground biomass prediction by Sentinel-1 multitemporal data in central Italy with integration of ALOS2 and Sentinel-2 data. *Journal of Applied Remote Sensing*, *12*, 1. https://doi.org/10.1117/1.JRS.12.016008

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., & Ceschia, E. (2017). Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. *Remote Sensing of Environment*, *199*, 415–426. https://doi.org/10.1016/j.rse.2017.07.015

Vermote, E. F., Tanre, D., Deuze, J. L., Herman, M., & Morcette, J.-J. (1997). Second simulation of the satellite signal in the solar Spectrum, 6S: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, *35*, 675–686. https://doi.org/10.1109/36.581987

Wang, G., Liu, S., Liu, T., Fu, Z., Yu, J., & Xue, B. (2019). Modelling aboveground biomass based on vegetation indexes: A modified approach for biomass estimation in semi-arid grasslands. *International Journal of Remote Sensing*, *40*, 3835–3854. https://doi.org/10.1080/01431161.2018.1553319

Wang, J., Brown, D., & Chen, J. (2013). Drivers of the dynamics in net primary productivity across ecological zones on the Mongolian plateau. *Landscape Ecology*, *28*, 725–739. https://doi.org/10.1007/s10980-013-9865-1

Wang, J., Xiao, X., Bajgain, R., Starks, P., Steiner, J., Doughty, R. B., & Chang, Q. (2019). Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *154*, 189–201. https://doi.org/10.1016/j.isprsjprs.2019.06.007

Wu, C., Shen, H., Shen, A., Deng, J., Gan, M., Zhu, J., Xu, H., & Wang, K. (2016). Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery. *Journal of Applied Remote Sensing*, *10*, 035010. https://doi.org/10.1117/1.JRS.10.035010

Zandler, H., Brenning, A., & Samimi, C. (2015). Potential of space-borne hyperspectral data for biomass quantification in an arid environment: Advantages and limitations. *Remote Sensing*, *7*, 4565–4580. https://doi.org/10.3390/rs70404565

Zhang, G., Kang, Y., Han, G., & Sakurai, K. (2010). Effect of climate change over the past half century on the distribution, extent and NPP of ecosystems of Inner Mongolia. *Global Change Biology*, *17*, 377–389. https://doi.org/10.1111/j.1365-2486.2010.02237.x

Zhang, Y., Kun, W., Zhang, J., Zhang, F., Xiao, H., Wang, F., Zhou, J., Song, Y., & Peng, L. (2021). Estimating rainfall with multi-resource data over East Asia based on machine learning. *Remote Sensing*, *13*, 3332. https://doi.org/10.3390/rs13163332

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ji, S., Dashpurev, B., Phan, T. N., Dorj, M., Jäschke, Y., & Lehnert, L. (2024). Above-ground biomass retrieval with multi-source data: Prediction and applicability analysis in Eastern Mongolia. *Land Degradation & Development*, *35*(9), 2982–2992. https://doi.org/10.1002/ldr.5109