Landsc Ecol (2023) 38:1463–1479
https://doi.org/10.1007/s10980-023-01613-1

RESEARCH ARTICLE

Biophysical drivers for predicting the distribution and abundance of invasive yellow sweetclover in the Northern Great Plains

Sakshi Saraf · Ranjeet John · Reza Goljani Amirkhiz · Venkatesh Kolluru · Khushboo Jain · Matthew Rigge · Vincenzo Giannico · Stephen Boyte · Jiquan Chen · Geoffrey Henebry · Meghann Jarchow · Raffaele Lafortezza

Received: 1 November 2022 / Accepted: 15 February 2023 / Published online: 25 March 2023

Abstract

Context Yellow sweetclover (*Melilotus officinalis*; YSC) is an invasive biennial legume that bloomed across the Northern Great Plains in 2018–2019 in response to above-average precipitation. YSC can increase nitrogen (N) levels and potentially cause substantial changes in the composition of native plant species communities. There is little knowledge of the spatiotemporal variability and conditions causing substantial widespread blooms of YSC across western South Dakota (SD).

Objectives We aimed to develop a generalized prediction model to predict the relative abundance of YSC in suitable habitats across rangelands of western South Dakota for 2019. Our research questions are: (1) What is the spatial extent of YSC across western South Dakota? (2) Which model can accurately predict the habitat and percent cover of YSC? and (3) What significant biophysical drivers affect its presence across western South Dakota?

Methods We trained machine learning models with *in situ* data (2016–2021), Sentinel 2A-derived surface reflectance and indices (10 m, 20 m) and site-specific variables of climate, topography, and edaphic factors to optimize model performance.

Results We identified moisture proxies (Shortwave Infrared reflectance and variability in Tasseled Cap Wetness) as the important predictors to explain the YSC presence. Land Surface Water Index and variability in summer temperature were the top predictors in explaining the YSC abundance. We demonstrated how machine learning algorithms could help generate valuable information on the spatial distribution of this invasive plant. We delineated major YSC

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10980-023-01613-1.

S. Saraf (✉) · R. John · R. Goljani Amirkhiz
Department of Biology, University of South Dakota, Vermillion, SD 57069, USA
e-mail: Sakshi.Saraf@coyotes.usd.edu

R. John · V. Kolluru · K. Jain · M. Jarchow
Department of Sustainability and Environment, University of South Dakota, Vermillion, SD 57069, USA

M. Rigge · S. Boyte
Center for Earth Resources Observation & Science, U. S. Geological Survey, Sioux Falls, SD 57198, USA

V. Giannico · R. Lafortezza
Department of Soil, Plant and Food Sciences (DISSPA), University of Bari "Aldo Moro", Via Amendola 165/A, 70126 Bari, Italy

J. Chen · G. Henebry
Department of Geography, Environment, and Spatial Sciences, and Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI 48823, USA

J. Chen
Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI 48823, USA

hotspots in Butte, Pennington, and Corson Counties of South Dakota. The floodplains of major rivers, including White and Bad Rivers, and areas around Badlands National Park also showed a higher occurrence probability and cover percentage.
Conclusions These prediction maps could aid land managers in devising management strategies for the regions that are prone to YSC outbreaks. The management workflow can also serve as a prototype for mapping other invasive plant species in similar regions.

Keywords Abundance · Habitat suitability model · Ensemble model · Northern Great Plains · Sentinel 2 · Plant invasion

Introduction

Invasive species have directly contributed to the decline of 49% of endangered and threatened species in the United States (Dueñas et al. 2018). Invasive species tend to reproduce rapidly and out-compete native species for food, water, and space, causing devastating effects on native biota (Northrup et al. 2019). These invasive species are altering ecosystem processes and functions, resulting in severe environmental damage and economic losses of ~ $120 billion annually in the United States alone (Fantle-Lepczyk et al. 2022). Grassland conversion to croplands and degradation due to overgrazing have contributed to increased soil erosion, diminished water quality, and the introduction of invasive plant species (Wright and Wimberly 2013).

Yellow sweetclover (*Melilotus officinalis*; YSC) is an invasive, semi-cryptophyte, annual-biennial, leguminous forb native to Eurasia. YSC was initially planted in the United States for bee habitat and to prevent roadside erosion, but was later utilized for livestock and wildlife forage (Gucker 2009). Changes in agricultural and soil conservation practices led to the establishment of YSC, and YSC acreage increased by 50% in South Dakota from 1927 to 1930 (Crosby and Kephart 1939). Seeding YSC was one of eight treatments applied to the dense clayey range sites of South Dakota in the spring of 1962 to increase productivity and accelerate the recovery of rangelands depleted by drought and overgrazing (Nichols and Johnson 1969). However, many positive traits associated with YSC, such as drought tolerance, high biomass, and capacity to germinate under various environmental conditions, indicate invasiveness and facilitate its dominance over native plant communities. The seeds of YSC are readily germinable and water-impermeable, and they remain viable in the soil seed bank for at least 20 years (Nelson 2018). YSC has very little top growth in its first year and puts its resources into establishing a deep root system and underground carbohydrate storage. In the second year, it thrives aboveground and blooms with yellow-colored flower racemes. If enough moisture is available for consecutive years, YSC possesses weedy attributes, i.e., it reproduces in large numbers and will remain persistent. YSC can sprout and survive in various environmental conditions (Ghaderi-Far et al. 2010), making it one of the highly invasive species in the rangelands of the Great Plains.

YSC is unique in its effect on nutrient flow and other ecosystem functioning. YSC, as a frontier invasive, has the potential to facilitate the establishment of invasive species through its substantial nitrogen (N) fixing and accumulation ability (84 kg N/ha/yr) across the Northern Great Plains, which are historically low-N systems (Lesica and DeLuca 2000; Van Riper and Larson 2009). Because native species are often better adapted to lower N levels, N-fixing plants (e.g.,YSC) provide a competitive advantage to invasive plant species and displace native perennials, resulting in a loss of diversity (Darbyshire and Small 2018). Wolf et al. (2003) found an increase in the number of exotic species and a concomitant decrease in native species when YSC colonized the montane grasslands of Colorado. YSC degraded the native grasslands and reduced biodiversity by covering and shading native plant species, changing the structure of grasslands and likely affecting wildlife population dynamics and behavior (Howard 2022).

YSC is an opportunistic plant that spreads quickly if favorable growth conditions are provided. Notably, South Dakota has had ~ 22% more 50 mm rain events since 1990 than the long-term average (Global Historical Climatology Network-Daily data for 30 stations from 1900 to 2020) (Frankson et al. 2022). The onset of these spring rain events drives YSC germination and helps it to thrive in the two subsequent growing seasons (Turkington and Cavers 1979; Vermeire and Rinella 2020). Given the projected warmer and wetter conditions in the Northern Great Plains (Hoell et al. 2020, 2021), the abundance of invasive species (e.g., YSC) are expected to increase in semi-arid ecosystems (Dukes and Mooney 1999; Frankson et al. 2022). Consecutive years of above-average precipitation in 2018–2019, were associated with widespread

and abundant YSC blooms (S.I. Figure 1). Historically, such years with YSC blooms are referred to as 'sweetclover years' (Gucker 2009).

YSC is drought-tolerant, mostly found in habitats associated with disturbances and dominated by other exotic species. YSC has the potential for establishing large biomass growing as tall as 2.4 m in its biennial life cycle, and providing habitat and nectar for honeybees (Gucker 2009). YSC is commonly observed in riparian areas, meadows, roadsides, railway embankments, and riverbanks. The wide distribution of YSC implies its tolerance to different environmental conditions. YSC can also be found in thick clayey-loam soil, and moist to slightly dry conditions. However, the mature plant size varies greatly depending on the soil's quality and the availability of moisture (Gucker 2009).

Although concern is growing regarding the YSC invasion of rangelands, there are several data gaps regarding its drivers and extent (Van Riper and Larson 2009). To the best of our knowledge, there are no studies regarding the distribution of YSC in South Dakota and the Northern Great Plains. We do not know the forcing mechanisms for the large and widespread blooms of invasive exotics like YSC (Merow et al. 2017). To fill this knowledge gap, we examine the contribution of biophysical factors to the distribution and extent of YSC using a suite of machine learning algorithms such as random forest, cubist, generalized additive model and extreme gradient boosting. Spatial modeling has been considered a promising approach for predicting the risk of invasion. Spatial invasion patterns can be predicted by linking the abundance and habitat of YSC to spatially explicit predictor variables. Climatic, topographic, or edaphic variables related to an increased probability of invasive species can be used to locate hotspots that could be targeted for monitoring and management efforts. Hence, YSC distribution maps at a broad spatial scale and a high resolution would be integral to identifying drivers, enabling targeted monitoring and management of YSC. Distribution maps can also help advance our knowledge of changes in the Northern Great Plains ecosystem's structure and function (Rigge et al. 2019).

We developed a generalized prediction model to predict the relative abundance of YSC in suitable habitats across rangelands of western South Dakota for 2019 (Saraf et al. 2023) https://doi.org/10.5066/P9X08W4T. We asked the following questions: 1.) What is the spatial extent of YSC across western South Dakota? 2.) Which model can accurately predict the habitat and percent cover of invasive YSC? and 3.) What environmental drivers affect its presence across western South Dakota? We hypothesized (**H1**) that YSC blooms are driven by a combination of wetter than average climate and topographic position (i.e., low-lying areas). We also hypothesized that YSC distribution is explained by a combination of (**H2**) heavy clay to loamy soils with alkaline to slightly acidic soil and above average soil moisture (Nichols and Johnson 1969; Turkington and Cavers 1979), and (**H3**) proximity to roads and stream networks, which facilitate dispersion.

Methods

Study area

We synthesized vegetation cover samples across the Upper Missouri River Basin (UMRB), which covers a large section of the Northern Great Plains. The UMRB covers 746,660 km^2 in six states: Montana, Nebraska, Wyoming, Minnesota, North Dakota, and South Dakota. We selected the UMRB, rather than just western South Dakota, for these samples to maximize the number of relevant observations and because the UMRB has experienced high variability in precipitation from 2016 to 2021, with flooding in 2019 and extreme drought in 2021 (Flanagan et al. 2020; Hoell et al. 2020, 2021; Frankson et al. 2022). Western South Dakota has an area of ~ 107,962 km^2, and covers approximately one-seventh of the UMRB (Fig. 1). This region is characterized by a continental climate with cold winters and warm summers. Our study focuses on western SD (west of the Missouri River), which experiences a semi-arid climate with high interannual variability in precipitation, averaging about 300–400 mm annually (Fig. 2). However, this region experienced above-average precipitation during 2019 (S.I. Figure 1). This area comprises old Mesozoic soils such as eroded clay, shale, and sandstone. The landscape consists of rolling hills, eroded stream valleys, and the Black Hills, with the primary land use being rangeland (Menéndez et al. 2020). Rangeland in western South Dakota is dominated by cool-season grasses such as western wheatgrass (*Pascopyrum smithii*), green needlegrass (*Nassella viridula*), smooth brome (*Bromus inermis*), and Kentucky bluegrass (*Poa*

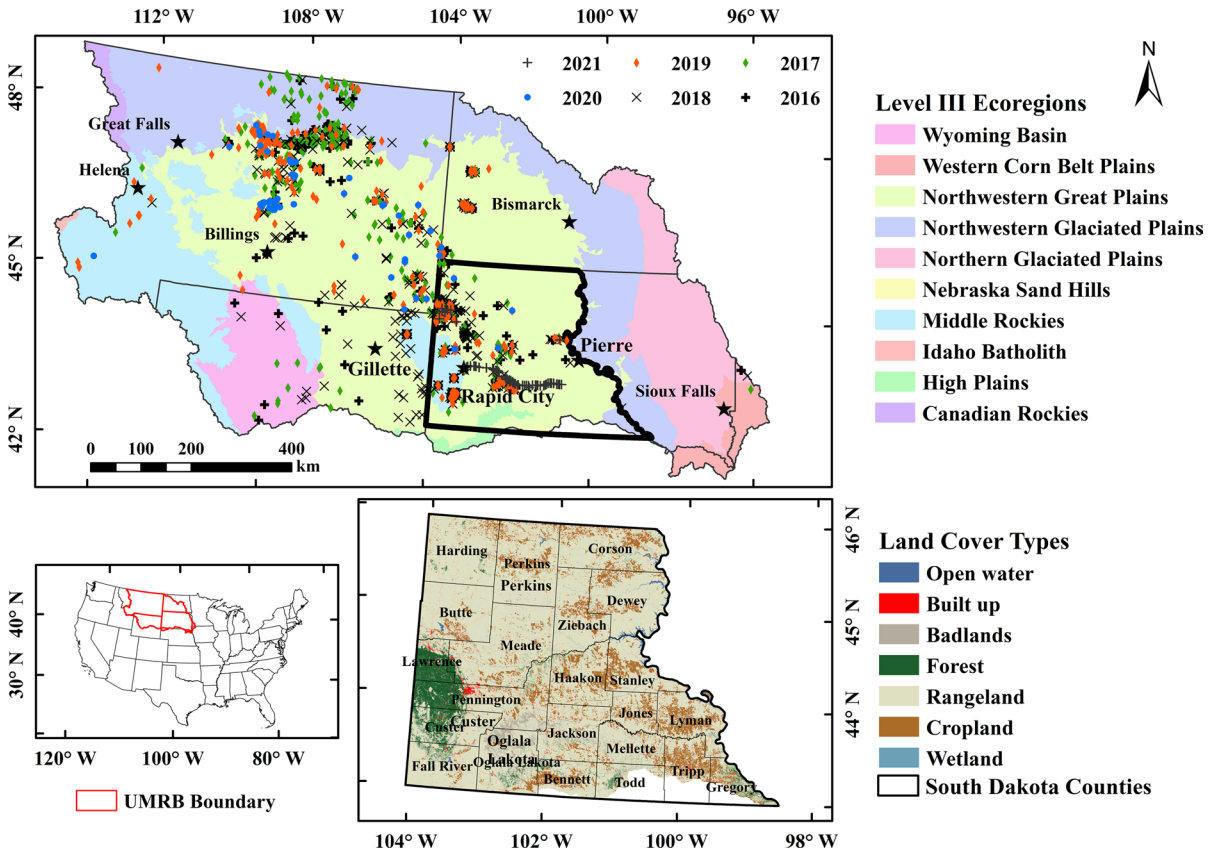

Fig. 1 Locations of field data collections (n = 2787) across the years 2016–2021, overlaid on the Level III ecoregions (U.S. Environmental Protection Agency –National Health and Environmental Effects Research Laboratory) of the Upper Missouri River Basin (top), and land cover map of western South Dakota overlaid with county boundaries (bottom right)

pratensis). This region also includes warm-season grasses, namely buffalo grass (*Bouteloua dactyloides*), blue grama (*Bouteloua gracilis*), little bluestem (*Schizachyrium scoparium*), switchgrass (*Panicum virgatum*), and forbs including western wallflower (*Erysimum capitatum*), Canada thistle (*Cirsium arvense*), leafy spurge (*Euphorbia esula*), and purple prairie clover (*Dalea purpurea*) (Rigge et al. 2013).

Data sources

Field measurements and sample collection

We aim to develop a model for YSC at 10 m spatial resolution. We compiled YSC percent cover and presence data from several databases of federal, state, and non-governmental organizations across the UMRB (S.I. Table 1). The state and source distribution of sample points are provided in S.I. Table 2, and the frequency distribution of YSC cover in these samples is shown in S.I. Figure 2. We collected 2787 sample points from different sources during 2016–2021 (S.I. Table 2). From these samples, we compiled 1612 percent cover samples, and 1438 presence-absence (binary) samples. Duplicates from different sources lying within the same pixel (10 m) and from the same year were removed. Observed YSC samples with less than 10% cover were discarded in the presence-absence model because we assumed canopy cover of < 10% would not be detectable in the 10 m spatial resolution imagery. For 2021, we conducted field measurements during peak summer months (June-July–August) across different biomes throughout western South Dakota (S.I. Table 2). We used a conventional plot-based quadrat method widely applied in the grassland ecology community

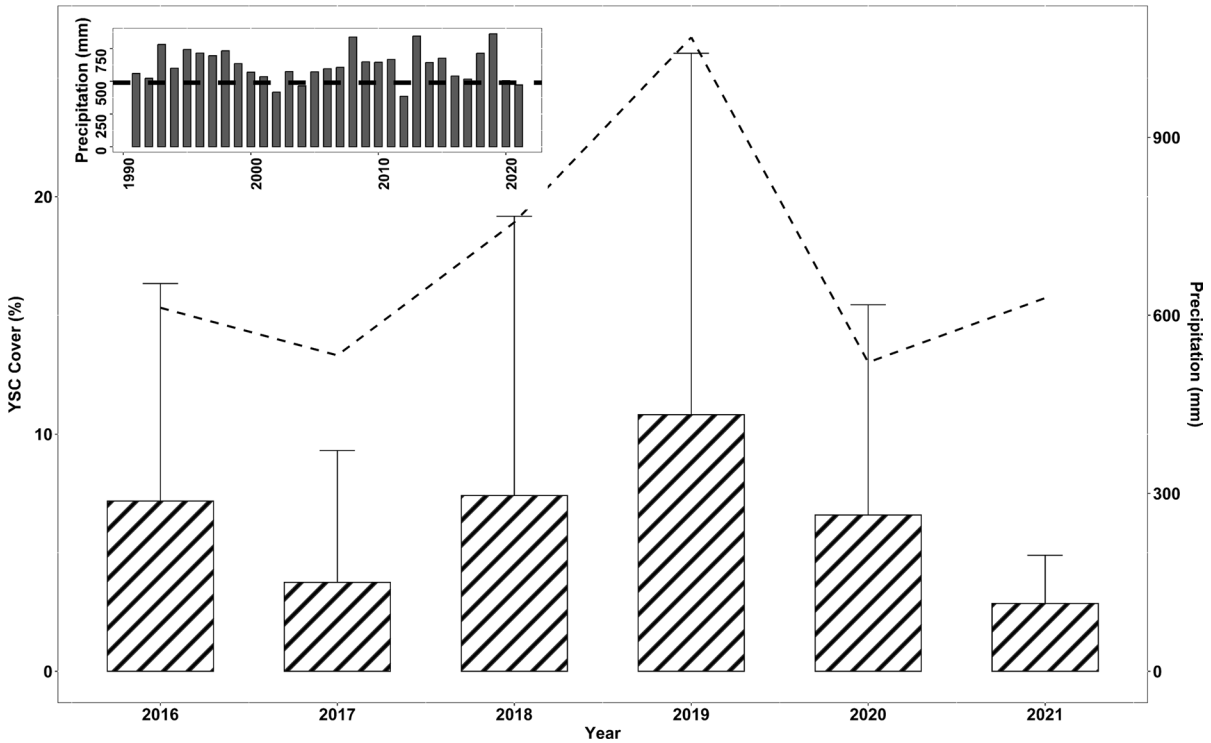

Fig. 2 Changes in the mean (standard deviation) of yellow sweetclover (YSC) percent canopy cover and the mean annual precipitation (dashed line) for western South Dakota for 2016–2021. The inset figure shows long-term mean annual precipitation (1991-2021), with the dashed line representing the 30-year average precipitation for the study region. Source details for YSC percent cover is mentioned in S.I. Table 1 and Precipitation was obtained from Daymet Data (S.I. Table 4)

to estimate canopy cover (John et al 2018) . In each 30×30 m plot, we randomly placed a 0.5×0.5 m quadrat at three locations. The quadrat was marked with a 5×5 cm grid. We calculated the percent cover of YSC by counting the number of YSC occupied grid cells.

Climatic and terrain variables

We obtained long-term daily (1991–2020) precipitation and temperature data at 1 km from the Daymet dataset (https://daymet.ornl.gov/) (Thornton et al. 2022). We computed annual (mean annual precipitation – MAP, mean annual temperature – MAT) and seasonal composites of precipitation (spring [March, April,May] – P_MAM and summer [June, July, August] – P_JJA) and temperature (T_MAM and T_JJA) using the Google Earth Engine (GEE). We also acquired normalized difference snow index (NDSI) data derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) MOD10A1 snow cover dataset, which has a spatial resolution of 500 m, to characterize the maximum percent snow cover in the region (Riggs et al. 2015). We also acquired snow depth from maximum value composites of monthly means of the ERA5-Land dataset processed by the European Centre for Medium-Range Weather Forecasts (ECMWF) at 11 km spatial resolution (Muñoz-Sabater et al. 2021). In addition, we used the National Elevation Dataset (10 m resolution) for terrain features to characterize elevation over the UMRB. We also obtained first-order derivatives such as slope, terrain roughness index and terrain wetness index from the National Elevation Dataset using the NASA Earthdata portal (Gesch et al. 2002, 2014).

Other biophysical variables

We used GEE to obtain cloud-free, or maximum 10% threshold cloud cover high-resolution Sentinel-2A surface reflectance images (Gascon et al. 2017) (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR). We derived several vegetation indices (VI) from Sentinel-2A: normalized difference vegetation index—(NDVI), green NDVI—(GNDVI), normalized difference water index—(NDWI), soil-adjusted vegetation index—(SAVI), and land surface water index—(LSWI). These are proxies for vegetation cover (NDVI, SAVI), greenness (GNDVI), water bodies (NDWI), and plant water content (LSWI). We also derived tasseled cap brightness (TCB), greenness (TCG), and wetness (TCW) indices from the 10 m Sentinel 2A bands, which serve as a proxy of albedo, vegetation, and moisture, respectively, to reduce dimensionality and the number of variables in rule-based modeling (Shi and Xu 2019). The Sentinel 2A-derived bands, VI and tasseled caps were computed using maximum value composites for the peak growing season (June–August) for each year from 2016 to 2021. Source details of all the VI's and tasseled caps used in this study are provided in S.I. Table 3. For soil textural and physical properties, we obtained soil pH, sand, silt, and clay from the Polaris database (Chaney et al. 2019), available at 30 m spatial resolution, developed by the National Cooperative Soil Survey under the US Department of Agriculture-Natural Resources Conservation Service. Land cover land use types and proximity to roads were derived from National Land Cover 2019 database (NLCD 2019) products available at 30 m resolution (Dewitz 2021). Lastly, the distance to streams product was derived from the National Hydrography Dataset developed by the U.S. Geological Survey National Geospatial Program (https://www.usgs.gov/national-hydrography). All data sets were resampled to 10 m spatial resolution and projected in Albers Equal Area projection and WGS 84 datum. We calculated standard deviations of all the variables derived from the monthly or annual composites to examine whether the variability of independent datasets explained the abundance of YSC. Altogether, we curated a database with 48 independent variables (S.I. Table 4).

The observed samples were segregated into binary presence-absence data for habitat suitability models and percent cover data for machine learning models. We used the presence-absence dataset to determine the best modeling approach for developing a habitat suitability map for YSC. In addition, we compared multiple models to determine the best approach to predicting the relative percent cover in the suitable habitats of YSC. The percent cover map was clipped

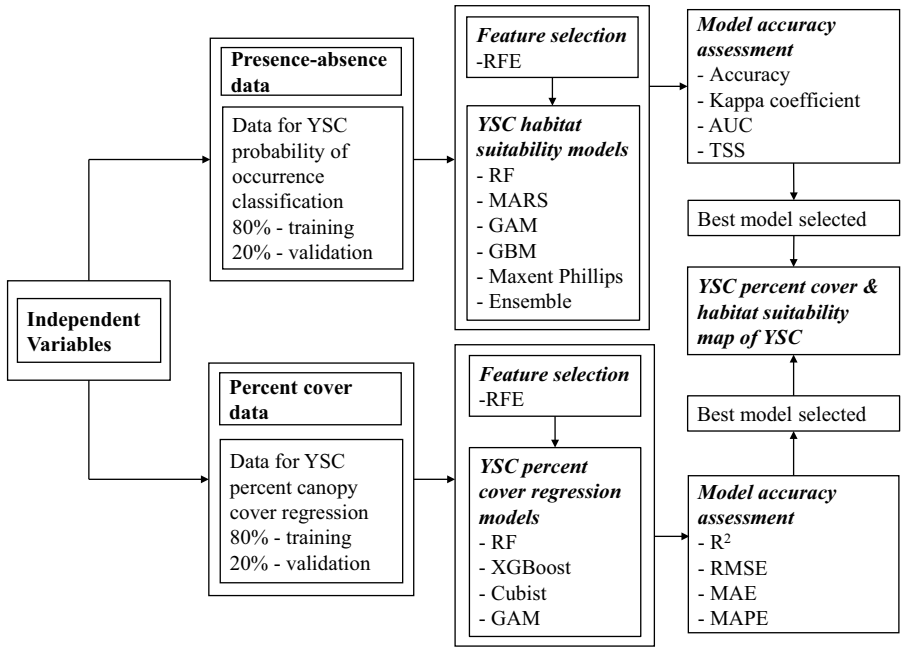

Fig. 3 Schematic workflow for predicting habitat suitability of yellow sweetclover (YSC) and its canopy cover (%) distribution across western South Dakota. *Note—*RFE* Recursive Feature Elimination, *RF* Random Forest, *MARS* Multivariate Adaptive Regression Splines, *GAM* Generalized Additive Models, *GBM* Generalized Boosting Model, *XGBoost* eXtreme Gradient Boosting model, *AUC* Area Under Curve, *TSS* True Skill Statistic, R^2 Coefficient of Determination, *R* Correlation Coefficient, *RMSE* Root Mean square error, *MAE* Mean Absolute Error, *MAPE* Mean Absolute Percentage Error

to the suitable habitats of YSC in the rangelands of western South Dakota (Fig. 1). The workflow to determine the suitable habitat of YSC and its percent cover distribution across rangelands of western South Dakota for 2019 is illustrated in Fig. 3.

For both habitat suitability and percent cover prediction models, we used a Pearson correlation coefficient (R) threshold of 0.7 (Green 1979; Domisch et al. 2013) to remove highly correlated variables. We also used the Recursive Feature Elimination (RFE) method to select the best subset of independent variables to help determine the suitable habitat and percent cover of YSC. The RFE (Guyon et al. 2002) selects all predictor variables and determines the relevance of each predictor using the backward selection technique. We re-iterated the model to discard all the least significant predictors based on significance scores. The best subset predictors from the RFE were used to train the machine learning algorithms (Guyon et al. 2002).

For both habitat suitability and percent cover prediction models, the samples were split into 80% for training and 20% for testing, using the bootstrapping method (including replacement). We used repeated cross-validation (CV) to optimize the hyperparameters (subset size) and select the best predictors based on variable importance ranking (Kuhn 2015) (S.I. Table 5). We implemented hyperparameter tuning and CV techniques to minimize errors and improve the model accuracy (Allouche et al. 2006).

Habitat suitability model

We tested five algorithms, namely the generalized additive model—(GAM; Hastie and Tibshirani 1987), generalized boosting model—(GBM; De'ath 2007), multivariate adaptive regression splines – (MARS; Friedman 1991), random forests—(RF; Breiman et al. 1984) and maximum entropy (Maxent Phillips) models (Phillips et al. 2006) to develop the habitat suitability model. These were all produced using the R statistical package *Biomod-2* (Thuiller et al. 2016). We implemented 10 evaluations with a threefold CV to evaluate the performance of individual models. The hyperparameters and CV technique details are described in S.I. Table 5. The average of the best three evaluated models was considered to develop the YSC prediction map.

We used accuracy and kappa coefficient to assess how well a model can predict and provide the proportion of correct classifications based on the training data (Thuiller 2003). We also implemented True Skill Statistic (TSS) and area under the receiver operating curve (ROC), also known as area under the curve (AUC), to evaluate specificity and sensitivity due to the imbalance between the presence and absence samples used in the model (Allouche et al. 2006; Amirkhiz et al. 2021).

The BIOMOD2 package's "random shuffle" approach was used to determine the importance of each variable (Thuiller et al. 2016). Three permutations per each variable were implemented to estimate the variable importance. We developed a response curve for each variable using the best-evaluated individual habitat suitability model to understand the influence of environmental factors on YSC. To create response curves for each covariate, we plotted the mean predicted probability of occurrence models, including the covariate, against the corresponding values of the covariate (where the other covariates in the model were held constant at their median value).

The best evaluated model for each individual algorithm was included in the ensemble-model building. The ensemble process within the BIOMOD2 framework, produced ensemble maps based on weighted means. All four metrics(Accuracy, Kappa, TSS and AUC) were used to evaluate the ensemble model. We used AUC score-based threshold to derive a binary map from the probability of occurrence as it was considered a standard to assess the performance of ordinal scoring models such as logistic regression (Allouche et al. 2006). The non-rangeland areas were masked out before implementing the habitat suitability model to emphasize the habitat of YSC in the rangelands of western South Dakota.

Estimating canopy cover (%)

We used 1612 samples from the YSC canopy cover dataset to develop four non-parametric machine learning-based regression models, namely the RF (Liaw and Wiener 2002; Kolluru and Kolluru 2021; Das et al. 2022), Cubist (John et al. 2018; Kuhn et al. 2023), extreme gradient boosting model – XGBoost (Abdullah et al. 2019; Chen et al. 2015) and GAM (Hastie 2015; Bera et al. 2021). Previous studies have

Table 1 Description of independent variables selected for estimating the probability of occurrence (presence-absence) and canopy cover (%) for yellow sweetclover (YSC)

S.no	Independent variables	Codes	Presence-absence (16)	Percent canopy (14)
1	Spring precipitation	P_MAM	*	*
2	Summer precipitation	P_JJA	*	
3	Spring temperature	T_MAM	*	
4	Summer temperature	T_JJA		*
5	Spring precipitation (standard deviation)	P_MAM_{sd}	*	*
6	Summer precipitation (standard deviation)	P_JJA_{sd}	*	
7	Summer temperature (standard deviation)	T_JJA_{sd}	*	*
8	Elevation	Elevation	*	*
9	Snow depth (standard deviation)	$Sdepth_{sd}$	*	*
10	Normalized difference snow index	NDSI		*
11	Standard deviation of NDSI	$NDSI_{sd}$		*
12	Land surface water index	LSWI	*	*
13	Normalized difference water index	NDWI	*	*
14	Sentinel 2A-Near infrared	S2A-NIR	*	*
15	Sentinel 2A-Shortwave infrared 1	S2A-SWIR1	*	*
16	Tasseled cap wetness (standard deviation)	TCW_{sd}	*	*
17	Soil pH	Soil_pH	*	
18	Percent silt	Silt	*	*
19	Percent clay	Clay	*	

*Indicates the variables used in the classification and regression models

shown that training a regression-tree model with various random ecological situations over multiple years can produce robust models with minimal extrapolation (Jacques et al. 2014; John et al. 2018). All variables were centered and scaled before the development of prediction models. All model hyperparameters were tuned; repeated CV was implemented to improve model fit and accuracy.

We evaluated the model performance using mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R^2) metrics during the training and testing phases. The model with the best statistical metrics was selected to predict YSC canopy cover (%) across suitable habitat regions of western South Dakota.

Results

Habitat suitability model

We selected 22 out of 48 predictor variables based on a correlation (R) threshold of $< \pm 0.7$ (S.I. Figure 5). We further ran RFE and selected the 16 best predictor variables that could explain the spatial variability of YSC (Table 1 and S.I. Figures 3 and 4). The RFE analysis showed that the most important variables associated with YSC habitat suitability were S2A-SWIR1, elevation, and spring precipitation (P_MAM).

Habitat suitability model performance comparison

We used an ensemble model of five individual models (RF, GAM, GBM, Maxent Phillips and MARS) to obtain the probability occurrence of YSC. Among individual models, RF performed best and MARS least (accuracy values of 0.87 and 0.80, respectively) and the ensemble model outperformed each individual model's predictions. The ensemble model scored best in the evaluation metrics with Accuracy, Kappa, AUC and TSS values 0.89, 0.72,0.92 and 0.73 respectively (S.I. Table 6). We used the AUC score-based threshold value of the ensemble model to develop the binary maps (S.I. Table 7). The predicted probability of occurrence range varied from 0 to 0.9 for the study region (Fig. 4).

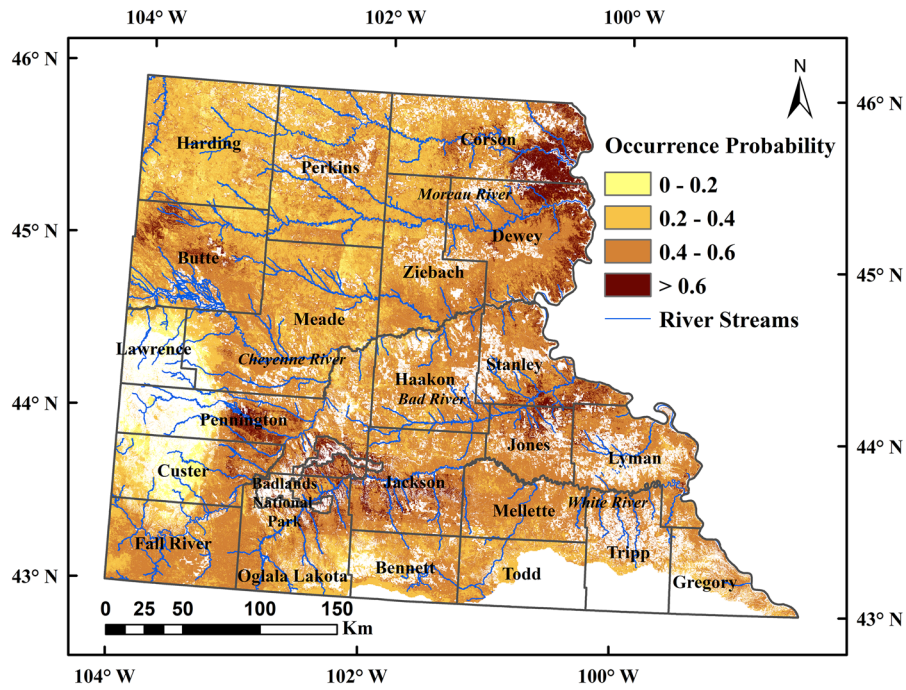

Fig. 4 Habitat suitability map for yellow sweetclover (YSC) determined by probability of occurrence based on an ensemble model across rangelands of western South Dakota

S2A-SWIR1, TCW_{sd}, S2A-NIR, and LSWI which are land surface property proxies of moisture and elevation were the most influential variables in determining YSC suitable habitat (S.I. Figure 6). The most important variables i.e., S2A-SWIR1, TCW_{sd} and LSWI, showed that higher soil moisture availability can explain the habitat suitability of YSC. Low range values of S2A-SWIR1 and TCW_{sd} and high range values of LSWI are known to represents high soil and canopy moisture content (Gao 1996; Xiao et al. 2002; Bajgain et al. 2015). Furthermore, S2A-NIR is often to represent healthy vegetation and is known to show a high correlation with the above ground biomass (Zhao et al. 2016). The higher range of S2A-NIR could perhaps represent higher above ground biomass of YSC. The response curves of seasonal precipitation variables (P_MAM and P_JJA) showed higher probability of occurrence even at below-average precipitation. The response curve for the topographic variables (elevation) showed higher probability of YSC occurrence in the low-lying areas. These results did support our first hypothesis that YSC presence can be explained by low-lying elevation. However, the first hypothesis that YSC requires above-average precipitation was not supported by seasonal precipitation. This could be attributed to coarser resolution data input.

In case of the second hypothesis, we found that clay and silt were the least important drivers of habitat suitability in the model. However, soil pH supported our hypothesis that YSC tends to occur more frequently in slightly acidic to alkaline soils. Although, the response curve for pH showed a higher probability of YSC occurrence in pH range of 6.5–8.0, the importance of this variable was only slightly higher than that of clay and silt. Furthermore, the response curves of S2A-SWIR1 supported our second hypothesis that YSC tends to grow with above-average soil moisture (Fig. 5 & S.I. Figure 7). Our third hypothesis regarding proximity factors (distance to roads and distance to streams) were not supported, as these variables were discarded based on RFE importance ranking. We used an optimum threshold of 0.44 to develop a binary presence-absence map from the probability of occurrence map (S.I. Table 7). Our final prediction map showed that more than half (42,722 km^2 out of 79,968 km^2) of western South Dakota rangelands provide a suitable habitat for YSC (S.I. Figure 8). The suitability of YSC across western South Dakota indicates that this species could exist in various environmental conditions. Major YSC hotspots were found in

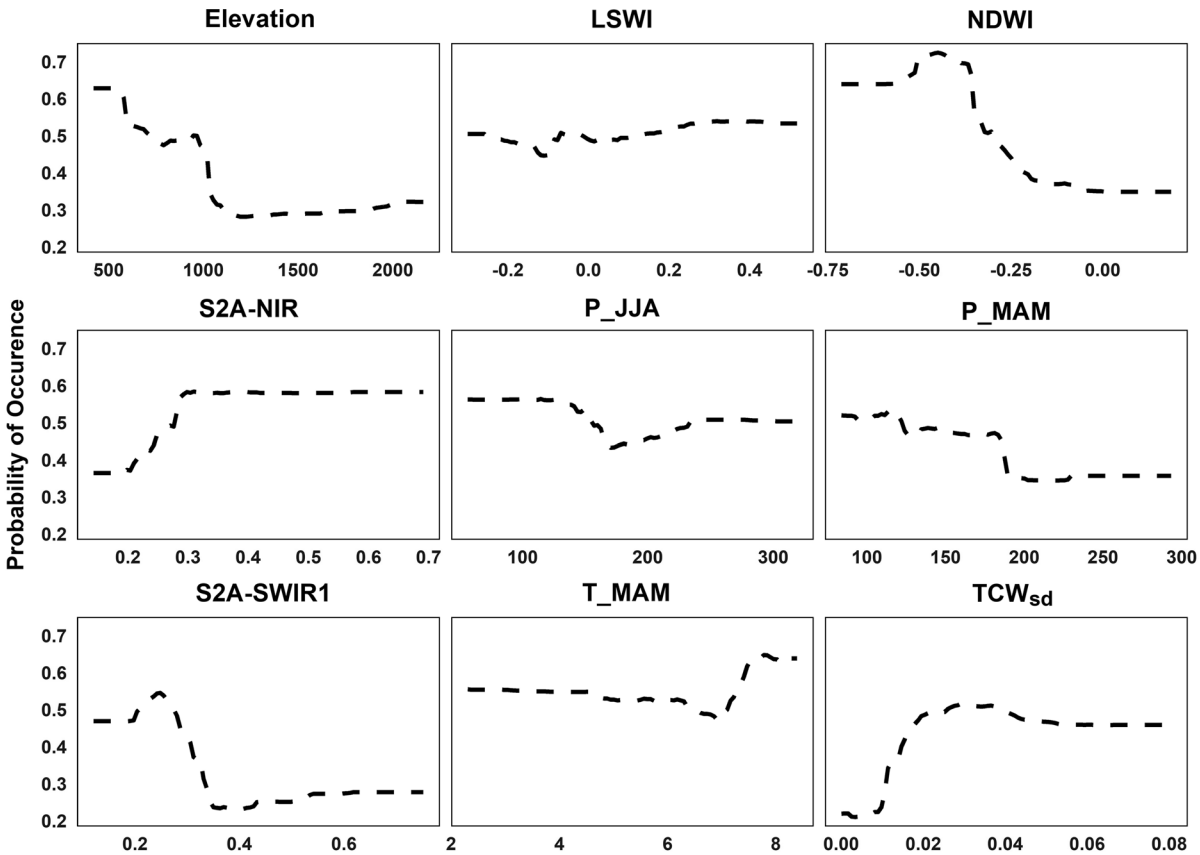

Fig. 5 Response curves for the top-ranked covariates of yellow sweetclover (YSC) derived from the random forest (best individual model) habitat suitability model. The response curves of remaining variables are shown in S.I. Figure 7. *Note—LSWI– Land surface water index, NDWI– Normalized difference water index, P_JJA—Summer precipitation, P_MAM—Spring precipitation, T_MAM—Spring temperature, TCW_{sd}—Variation in tasseled cap wetness

(i) portions of Butte, Pennington, Corson and Dewey Counties, (ii) areas around Badlands National Park, and (iii) some areas adjacent to floodplains of the White and Bad Rivers or their tributaries.

Estimated canopy cover and model performance

We selected 21 of 48 predictor variables based on a correlation (R) threshold of $< \pm 0.7$ (S.I. Figure 9). We further ran RFE and selected the 14 best variables for predicting YSC percent cover distribution (Table 1 and S.I. Figures 3 and 4). RFE analysis showed that the most important variables associated with YSC percent cover were LSWI, variation in summer temperature (T_JJA_{sd}), and S2A-SWIR1.

We implemented four regression models (i.e., GAM, RF, Cubist and XGBoost) using the *caret* package in R (Kuhn et al. 2020). Based on the training and testing evaluation metrics, the GAM model had the poorest performance, with an R^2 of 0.30 during the testing phase. XGBoost performed relatively well compared to GAM but exhibited overfitting of the data. XGBoost had an R^2 of 0.88 during training, but the metrics decreased to 0.46 during the testing phase. Cubist performed second best with an R^2 of 0.54 during the testing phase. RF yielded the best result with an R^2 of 0.55 and RMSE of 7.49 during the testing phase. In addition to R^2 and RMSE, we found that RF had an MAE and MAPE of 5.09 and 2.76, respectively (Table 2).

The RF model was chosen to estimate the percent cover distribution of YSC in western SD. The variable importance of the RF model showed that LSWI performed the best and was among the top influencing

Table 2 Statistical comparison of machine learning models for predicting canopy cover (%) of yellow sweetclover (YSC) along with their descending order of variable importance (See S.I. Table 4 for variable descriptions)

Model	R^2	R	RMSE	MAE	MAPE	Variable importance
GAM	0.30	0.54	9.13	6.50	3.21	Elevation, P_MAM, T_JJA, TCW_{sd}, S2A-SWIR1
Cubist	0.54	0.73	7.48	4.99	2.52	LSWI, Elevation, T_JJA_{sd}, S2A-SWIR1, P_MAM, TCW_{sd}
Random forest	0.55	0.74	7.49	5.09	2.76	LSWI, T_JJA_{sd}, S2A-SWIR1, TCW_{sd}, $NDSI_{sd}$, P_MAM
XGBoost	0.46	0.68	8.02	4.96	2.56	LSWI, S2A-NIR, S2A-SWIR1, Elevation, P_MAM, $NDSI_{sd}$

GAM Generalized Additive Models, *XGBoost* eXtreme Gradient Boosting, R^2 Coefficient of Determination, *R* Correlation coefficient *RMSE* Root Mean square error, *MAE* Mean Absolute Error, *MAPE* Mean Absolute Percentage Error

variables in other modeling approaches (S.I. Figure 10 and Table 2). Following LSWI, summer temperature variability (T_JJA_{sd}), S2A-SWIR1, and variability in tasseled cap wetness (TCW_{sd}) were among the other top important variables for both the RF and Cubist models. The response curves for the RF percent cover model are represented in S.I. Figure 11. The response curves for YSC % cover distribution showed increase in LSWI significantly increased YSC % cover range from 0 to 20% whereas other variables explained small variation in abundance ranging between 0 and 10%.

We also developed a spatial correlation map using habitat suitability and predicted canopy cover (%) distribution of YSC (S.I. Figure 12). Small sections of western counties such as Butte, Meade, Pennington, and Fall River showed negative correlation. Most of the study region showed a positive correlation between the probability of occurrence and percent cover distribution of YSC. The predicted canopy cover map was clipped to emphasize the suitable habitat of YSC in the rangelands of western South Dakota. The predicted YSC percent cover distribution had higher heterogeneity and varied along moisture gradients. Highest canopy cover

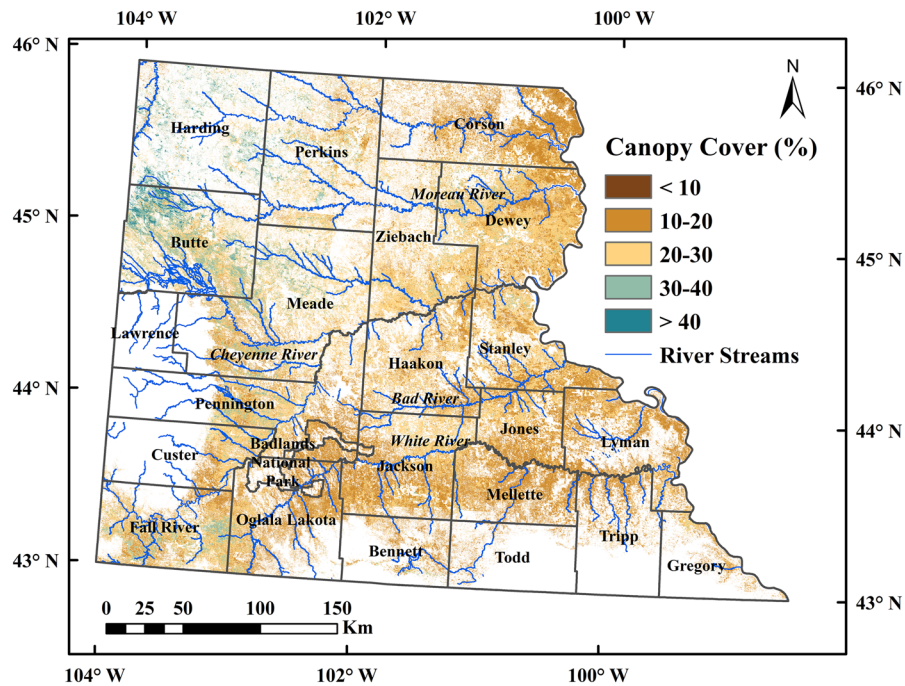

Fig. 6 Predicted canopy cover map of yellow sweetclover (YSC) based on random forest (best) model across rangelands of western South Dakota for 2019

distribution was found in parts of Butte County followed by Meade, Fall River, and Pennington Counties. The predicted percent cover in suitable areas of YSC ranged from 5 to 53% with relatively higher abundance in the western counties of western South Dakota (Fig. 6).

Discussion

We developed a framework to predict the spatial distribution and cover of YSC across western South Dakota. We gathered YSC samples across multiple years (2016–2021), including wet and dry years to account for spatial heterogeneity and temporal variation in the model. Sampling data over multiple growing seasons can help obtain a robust calibration model (John et al. 2018). We first developed a habitat suitability model and inspected response curves to understand the importance of various abiotic and biotic factors to YSC distribution. We then developed a prediction model of YSC canopy cover to determine the abundance of YSC. The abundance map can be used as an indicator of site conditions and as a surrogate index of habitat suitability. We tested different machine learning models, which varied in their performance (Tables 2 & S.I. Table 6). Different types of data, such as presence-absence, presence only, or percent cover data or the species-specific traits of habitat, may take precedence in local model selection (Stohlgren et al. 2010). Testing several predictive models and several forms of datasets prevents the bias of relying on a single model. This approach of comparing various predictive models adds substantial robustness and consistency to the performance of a specific predictive model (Kolluru et al. 2020).

Effect of environmental conditions

Climatic drivers

Precipitation is one of the key factors affecting grassland structure and function (Lauenroth et al. 2014). Previous studies have found an amplified response of invasive annuals to precipitation characteristics (Bradley and Mustard 2005). We found that spring precipitation (P_MAM) was a moderately significant climatic variable in predicting YSC presence. Generally, YSC emerges around spring (March or April) and grows approximately 10.2–30.5 cm tall, provided enough moisture is available during the growth period during the first year of its life cycle. Once the root develops, the plant becomes drought and cold-tolerant (Luo et al. 2016). The response curve results showed that low spring precipitation ranging between 75 and 175 mm may contribute to higher chances of YSC occurrence (Fig. 5). This could perhaps be owing to coarse-resolution precipitation data inputs or less sample points. The response curves may have under-represented the range of spring precipitation. However, the higher resolution land surface property variables (LSWI, S2A-SWIR1, TCW_{sd}) indicated that YSC distribution is explained by high moisture availability during its vegetative growth period. The higher range of moisture proxies with low season precipitation range may indicate that the growth of YSC requires accumulated moisture from two consecutive wet periods to show enhanced productivity. This effect of accumulated moisture could not be explained by seasonal precipitation variables (P_MAM and P_JJA) on their own but could have perhaps been explained by annual precipitation. However, annual precipitation was removed from this study due to high collinearity with other predictor variables. Further research would be useful to understand the interaction or combined effect of climate variables on YSC presence and abundance.

Temperature is a significant factor in various physiological processes in addition to precipitation (Gama et al. 2016). The periodicity of seed germination and spread of species are significantly influenced by temperature (Ghaderi-Far et al. 2010). Conn et al. (2011) showed that YSC could even survive the extreme weather in Alaska, with annual precipitation and temperature of 170 mm and − 3.3 °C, respectively. Germination rate generally rises linearly with temperature and drops rapidly at a higher temperature within a narrow range (Finch-Savage 2020). We found a slightly different trend in the response curve of spring temperature. Although YSC probability of occurrence gradually decreased until 7 °C, a slight increase in temperature (T_MAM) from 7 to 8 °C increased YSC probability of occurrence from 0.45 to 0.65 (Fig. 5).

Furthermore, we also found that variation in percent snow cover ($NDSI_{sd}$) in winters has a significant

effect on YSC abundance. Paudel and Andersen (2013) found a positive linear relationship between rangeland vegetation production and snow cover in the drier rangelands of the Trans Himalayan Region. Their findings indicated that although temperature is important, changes in snow cover and precipitation patterns play more important roles in snow-fed, drier regions for rangeland vegetation dynamics.

Effect of topographic drivers

Elevation gradient can affect and alter the seasonal development of various plant species. The response curve showed decreasing occurrence of YSC with increasing elevation. The probability of YSC occurrence dropped from 0.6 to 0.2 when the elevation increased beyond 1000 m (Fig. 5). YSC dominance is common at low to medium elevation near streambanks, meadows, or disturbed areas (Hansen et al. 1988). Previous studies have found that the YSC can grow even at higher elevations of above 2000 m (Stoa 1933). However, higher elevation sites can affect vegetative growth of YSC by affecting its flowering period, fruit development, and seed maturation. We also found that the YSC abundance increased moving from east to west across the elevation gradient (Fig. 6).

Effect of edaphic variables

We hypothesized that heavy clay to loamy soils might be influential in determining the presence of YSC. However, we found that clay and silt were the least important in the habitat suitability model whereas silt was the least important in predicting YSC abundance. This could be because of its ability to grow in various soil types. On the contrary, seeding of YSC on South Dakota's rangelands is successful where soils are as much as 65% clay (Nichols and Johnson 1969). YSC is also productive on silty loams to clayey loam soil with neutral to alkaline pH. We found that soil pH had little effect on YSC habitat suitability. The response curve indicated a pH range of 6.5–8.0 has the highest probability of YSC occurrence. Our results align with Turkington and Cavers (1979), who showed that YSC could thrive on soils with a pH > 6.5, whereas YSC is less competitive and fails at nitrogen fixation in acidic soils.

Distribution along roads and streams

YSC is often associated with disturbed lands and open sites. However, most of the undisturbed lands located farther from roads has already been affected by YSC in South Dakota. Proximity to roads and streams as two separate variables were discarded for both habitat suitability and percent cover model during variable selection. This could be due to fewer field samples collected near roads and streams to minimize bias. The percent cover and suitability map did not show any bias towards the roads or stream networks. The dispersion of sweetclover (yellow and white sweetclover) via roads or river streams were important factors in introducing and establishing exotic species (Tikka et al. 2000). Wurtz et al. (2010) conducted a case study and documented the distribution of sweetclover near river crossings. They suggest that sweetclover might have spread onto the floodplains from upstream roads, mines, and agricultural developments. Early surveys for exotic plant species showed that sweetclover had invaded the Stikine, Matanuska, and Nenana Rivers floodplains in Alaska (Conn et al. 2008). The dispersion of YSC via roads or highways had an important role in introducing and establishing exotic species. Tikka et al. (2000) observed that grassland plants leveraged roads and railway corridors for dispersion in central Finland. Contrary to not finding the distance to roads among the most important variables in this study, we found several patches of YSC alongside the highways and interstates of South Dakota during field visits. However, the length and size of these patches along the highways vary.

Effect of other biophysical variables

YSC is ubiquitous across the Northern Great Plains, but very little is known about its spatial distributions and/or the environmental factors that affect its growth (Van Riper and Larson 2009). S2A-SWIR1 was the top indicator for predicting YSC presence and the third most important indicator for estimating its canopy cover (S.I. Figure 6 & S.I. Figure 10). S2A-SWIR-1 is sensitive to the moisture content in soil and vegetation. S2A-SWIR1 reflectance decreases as the water

content increases in the soil. These characteristics make it helpful in distinguishing between dry and wet soils. YSC's response to S2A-SWIR1 strengthens the understanding that YSC either prefers to grow in moist areas or it has high canopy water content. YSC also showed an upward abundance trend with the increase in LSWI, which was the top variable for predicting percent canopy cover and the fourth most important for predicting the habitat of YSC. LSWI also includes the S2A-SWIR1 band and is sensitive to volumetric soil moisture and canopy water content (Gao 1996; Xiao et al. 2002; Bajgain et al. 2015). The response curve showed probabilility of YSC occurrence if the TCW_{sd} range varied upto 0.08 whereas the actual range of TCW_{sd} varied upto 0.5 (Fig. 5 & S.I. Figure 4). The chances of YSC occurrence are higher only if consistent moisture availability is present in the region.

Implications for the management of YSC

Monitoring and mapping of YSC can help examine its current distribution patterns. These methods can help managers anticipate the possible trajectory of a species and its distribution in the coming years (Howard 2022). YSC has many positive attributes, such as a positive economic impact on honey production, its usage as a forage crop, carbon sequestration and soil stabilization. However, YSC can also threaten native prairies because it can reproduce in large numbers and effectively compete for nutrients with desired native species (Van Riper and Larson 2009). Therefore, developing an operational and cost-effective monitoring system would be beneficial to understand the growth and spread of YSC in a specific region. The present study could help identify potential hotspot areas for implementing management strategies to stabilize or manage the level of YSC growth. A monitoring system could also help honey producers track their potential production for the upcoming years. Hence, mapping YSC in near-real time, at a broad spatial scale and high resolution, could help improve land management and policymaking decisions and provide better knowledge of YSC habitat and its distribution.

Limitations of the study and future scope

The use of regression tree approaches such as RF, XGBoost and Cubist in estimating the canopy cover (%) has some limitations that could be addressed in future research. First, a small dataset (n = 1612) and the skewness of the data were a major limitation when using models such as RF and Cubist. Larger datasets would ideally be analyzed to improve the model's precision and reliability. A more stratified sample collection from disturbed and undisturbed sites could help explain its spread pattern. Many rangeland plant species are indistinguishable in aerial or satellite imagery except in specific phenophases. YSC is a biennial species that grows over two years. It grows approximately 10.2–30.5 cm tall in the first year but does not bloom until the second year, making it harder to detect YSC in its first year using multi-spectral imagery at 10 m resolution. The data collected from various sources over the past years did not have information regarding the phenological growth stages of YSC (bloom/no-bloom). The absence of phenological growth information on YSC could lead to error propagation during response variable generation in geographic information system environments. We found all the top predictor variables belonged to high-resolution dataset. Subsampling of datasets due to unavailability of climate variables at fine resolution is still a major limitation for 10 m resolution small-scale studies.We used the POLARIS dataset despite availability of Soil Survey Geographic (SSURGO) database at 10 m resolution because POLARIS data overcomes the challenges of spatial details and incomplete spatial coverage (Chaney et al. 2019). In addition, dimensionality reduction using multi-collinearity and variable selection might eliminate important variables that can explain the species occurrence. Future studies could adapt and test more environmental factors and regression models in estimating YSC. The percent cover model evaluation metrics showed a good model fit, and the data strongly supported the classification of YSC suitable areas. These generalized models ideally would be validated with more observation samples in the forthcoming years. The present study can be considered a reference that can be improved with more observations, phenological stage datasets, environmental layers and resolution between temporal and spatial variation.

Conclusions

We predicted YSC habitat suitability and its percent cover across western South Dakota using machine learning algorithms trained with in situ field samples and biophysical drivers from different remote sensing datasets. Our study identified the most suitable drivers that could explain the variability in YSC presence and its percent cover through data dimensionality reduction techniques (i.e., correlation and recursive feature elimination). We compared different modeling techniques and used the best model to develop a classified YSC presence-absence and canopy cover (%) distribution map for YSC in 2019. We also examined the variable importance and their response curves to determine the suitable habitat of YSC across western South Dakota. Our research demonstrated how ensemble techniques for habitat suitability models and machine learning algorithms such as RF could help generate valuable information on the spatial distribution of an invasive rangeland plant species. These machine learning models helped explore the environmental constraints in distinguishing invasive YSC through satellite imagery. The Northern Great Plains ecosystem is vulnerable to invasive species, prolonged droughts, human disturbance, and over-grazing (Derner et al. 2018). Therefore, our prediction maps could be used to create a strategies that can be used by managers for regions vulnerable to YSC. Moreover, these results will support ranchers in implementing conservation plans to protect vulnerable species and support beekeepers in making the best productive use of these species. Mapping invasive species in the Northern Great Plains every year or two would be useful to monitor the spatiotemporal dynamics of grassland landscapes. Monitoring plant invasives can provide landscape ecologists with information on the spatiotemporal scales of change in grasslands, allowing them to make more effective decisions. This research can help in mapping other invasive species in the rangeland ecosystem.

Acknowledgements We thank all the institutes, organizations, and developers of the various datasets for making their products freely available. Thanks are extended to Dr. Marissa Ahlering and Alison Long (The Nature Conservancy), Dr. Amy Symstad, Dr. Aaron Johnston, and Todd Preston (U.S. Geological Survey), Jacob Dyer (South Dakota Game Fish, and Parks), Milton Haar (Badlands National Park and National Park Service Inventory & Monitoring Network), Carmen Drieling (Bureau of Land Management), and Braden Burkholder (Montana Natural Heritage Program), Bruce Wylie for their support and for freely sharing their data. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author contributions RJ and SS contributed to the study's conception and design. SS, VK and KJ performed data collection; SS performed the material preparation, analysis and prepared the first draft of the manuscript. All authors read and approved the manuscript.

Funding This study was supported by the South Dakota Board of Regents(SDBOR) Competitive Research Grant and Research Infrastructure Development (RDI), NSF RII Track-4:NSF2229746, NASA LCLUC Program-80NSSC20K0410, and NSF RII Track-2 FEC1632810.

Declarations

Competing interest The authors declare no competing interest.

References

Abdullah AYM, Masrur A, Gani Adnan MS et al (2019) Spatio-temporal patterns of land use/land cover change in the heterogeneous coastal region of Bangladesh between 1990 and 2017. Remote Sens. https://doi.org/10.3390/rs11070790

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43:1223–1232

Amirkhiz RG, Dixon MD, Palmer JS, Swanson DL (2021) Investigating niches and distribution of a rare species in a hierarchical framework: Virginia's Warbler (Leiothlypis virginiae) at its northeastern range limit. Landsc Ecol 36:1039–1054

Bajgain R, Xiao X, Wagle P et al (2015) Sensitivity analysis of vegetation indices to drought over two tallgrass prairie sites. ISPRS J Photogramm Remote Sens 108:151–160

Bera D, Das CN, Bera S (2021) Comparative performance of linear regression, polynomial regression and generalized additive model for canopy cover estimation in the dry deciduous forest of West Bengal. Remote Sens Appl Soc Environ 22:100502

Bradley BA, Mustard JF (2005) Identifying land cover variability distinct from land cover change: cheatgrass in the Great Basin. Remote Sens Environ 94:204–213

Breiman L, Friedman JH, Olshen RA, Stone C (1984) Classification and regression trees, 1st edn. Routledge, Oxfordshire

Chaney NW, Minasny B, Herman JD et al (2019) POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. Water Resour Res 55:2916–2938

Chen, T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Mitchell R, Cano I, Zhou, T (2015) Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4

Conn JS, Beattie KL, Shephard MA et al (2008) Alaska Melilotus invasions: distribution, origin, and susceptibility of plant communities. Arctic, Antarct Alp Res 40:298–308
Conn J, Werdin-Pfisterer N, Beattie K, Densmore R (2011) Ecology of invasive melilotus albus on Alaskan glacial river floodplains. Arctic, Antarct Alp Res 43:343–354
Crosby MA, Kephart LW (1939) Sweetclover in corn belt farming. US Department of Agriculture
Darbyshire S, Small E (2018) Are melilotus albus and M. officinalis conspecific? Genet Resour Crop Evol 65:1571–1580
Das P, Panda RM, Dash P et al (2022) Multi-decadal mapping and climate modelling indicates eastward rubber plantation expansion in India. Sustainability. https://doi.org/10.3390/su14137923
De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecology 88:243–251
Derner J, Briske D, Reeves M et al (2018) Vulnerability of grazing and confined livestock in the Northern Great Plains to projected mid- and late-twenty-first century climate. Clim Change 146:19–32
Dewitz J (2021) National Land Cover Database (NLCD) 2019 products. US Geol Surv Sioux Falls, SD, USA. https://doi.org/10.5066/P9KZCM54
Domisch S, Kuemmerlen M, Jähnig SC, Haase P (2013) Choice of study area and predictors affect habitat suitability projections, but not the performance of species distribution models of stream biota. Ecol Modell 257:1–10
Dueñas M-A, Ruffhead HJ, Wakefield NH et al (2018) The role played by invasive species in interactions with endangered and threatened species in the United States: a systematic review. Biodivers Conserv 27:3171–3183
Dukes JS, Mooney HA (1999) Does global change increase the success of biological invaders? Trends Ecol Evol 14:135–139
Fantle-Lepczyk JE, Haubrock PJ, Kramer AM et al (2022) Economic costs of biological invasions in the United States. Sci Total Environ 806:151318
Finch-Savage WE (2020) Influence of seed quality on crop establishment, growth, and yield. In: Gough RE, Basra AS (eds) Seed quality. CRC Press, Boca Raton, pp 361–384
Flanagan PX, Mahmood R, Umphlett NA et al (2020) A hydrometeorological assessment of the historic 2019 flood of Nebraska, Iowa, and South Dakota. Bull Am Meteorol Soc 101:E817–E829
Frankson R, Kunkel KE, Champion SM et al (2022) South Dakota State Climate Summary 2022. In: NOAA Tech. Rep. NESDIS 150-SD. NOAA/NESDIS, Silver Spring, MD, 5 pp. https://statesummaries.ncics.org/chapter/sd/. Accessed 1 July 2022
Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–141
Gama M, Crespo D, Dolbeth M, Anastácio P (2016) Predicting global habitat suitability for corbicula fluminea using species distribution models: the importance of different environmental datasets. Ecol Modell 319:163–169
Gao B (1996) NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens Environ 58:257–266
Gascon F, Bouzinac C, Thépaut O et al (2017) Copernicus Sentinel-2A calibration and products validation status. Remote Sens. https://doi.org/10.3390/rs9060584
Gesch DB, Oimoen MJ, Evans GA (2014) Accuracy Assessment of the U.S. Geological Survey National Elevation Dataset, and Comparison with Other Large-Area Elevation Datasets-SRTM and ASTER Open-File Report 2014–1008. US Geol Surv Open-File Rep 2014–1008 18. https://doi.org/10.3133/ofr20141008
Gesch D, Oimoen M, Greenlee S, Nelson C, Steuck M, Tyler D (2002) The national elevation dataset. Photogramm Eng Remote Sens 68(1):5–32
Ghaderi-Far F, Gherekhloo J, Alimagham M (2010) Influence of environmental factors on seed germination and seedling emergence of yellow sweet clover (Melilotus officinalis). Planta Daninha 28:463–469
Green RH (1979) Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York
Gucker CL (2009) Melilotus alba, M. officinalis, Fire Effec. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory
Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422
Hansen PL, Chadde SW, Pfister RD (1988) Riparian dominance types of Montana. Misc. Publ. No. 49. University of Montana, School of Forestry, Montana Forest and Conservation Experiment Station, Missoula, MT, pp 5660
Hastie T (2015) Package 'gam'. R package version, 90124–3
Hastie T, Tibshirani R (1987) Generalized additive models: Some applications. J Am Stat Assoc 82:371–386
Hoell A, Parker BA, Downey M et al (2020) Lessons learned from the 2017 flash drought across the U.S. Northern Great Plains and Canadian Prairies. Bull Am Meteorol Soc 101:E2171–E2185
Hoell A, Hoerling M, Eischeid J, Barsugli J (2021) Preconditions for extreme wet winters over the contiguous United States. Weather Clim Extrem 33:100333
Howard TM (2022) Invasive species in forests and Rangelands of the United States: a comprehensive science synthesis for the United States forest sector. Soc Nat Resour 35:338–339
Jacques DC, Kergoat L, Hiernaux P et al (2014) Monitoring dry vegetation masses in semi-arid areas with MODIS SWIR bands. Remote Sens Environ 153:40–49
John R, Chen J, Giannico V et al (2018) Grassland canopy cover and aboveground biomass in Mongolia and Inner Mongolia: spatiotemporal estimates and controlling factors. Remote Sens Environ 213:34–48. https://doi.org/10.1016/j.rse.2018.05.002
Kolluru V, Kolluru S (2021) Development and evaluation of pre and post integration techniques for enhancing drought predictions over India. Int J Climatol 41:4804–4824
Kolluru V, Kolluru S, Wagle N, Acharya TD (2020) Secondary precipitation estimate merging using machine learning: development and evaluation over Krishna River Basin, India. Remote Sens. https://doi.org/10.3390/rs12183013
Kuhn M (2015) A Short Introduction to the caret Package. R Found Stat Comput 1–10

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team RC (2020) Package 'caret'. R J 223(7)
Kuhn M, Weston S, Keefer C (2023) Package 'Cubist'. Rule- and Instance-Based Regression Modeling. R Package Version 0.4, 1. R Package Version 0.4, 1
Lauenroth WK, Schlaepfer DR, Bradford JB (2014) Ecohydrology of dry regions: storage versus pulse soil water dynamics. Ecosystems 17:1469–1479
Lesica PL, DeLuca TH (2000) Sweetclover: a potential problem for the Northern Great Plains. J Soil Water Conserv 55:259–261
Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22
Luo K, Jahufer MZZ, Wu F et al (2016) Genotypie variation in a breeding population of yellow sweet clover (Melilotus officinalis). Front Plant Sci 7:1–10
Menéndez P, Losada IJ, Torres-Ortega S et al (2020) The global flood protection benefits of Mangroves. Sci Rep 10:1–11
Merow C, Bois ST, Allen JM et al (2017) Climate change both facilitates and inhibits invasive plant ranges in New England. Proc Natl Acad Sci U S A 114:E3276–E3284
Muñoz-Sabater J, Dutra E, Agusti-Panareda A et al (2021) ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst Sci Data 13:4349–4383
Nelson SO (2018) RF Electrical Seed Treatment to Improve Germination. In: 2018 ASABE Annual International Meeting. p 1
Nichols JT, Johnson JR (1969) Range productivity as influenced by biennial sweetclover in Western South Dakota. J Range Manag 22:342
Northrup JM, Rivers JW, Yang Z, Betts MG (2019) Synergistic effects of climate and land-use change influence broad-scale avian population declines. Glob Chang Biol 25:1561–1575
Paudel KP, Andersen P (2013) Response of rangeland vegetation to snow cover dynamics in Nepal Trans Himalaya. Clim Change 117:149–162
Phillips SB, Aneja VP, Kang D, Arya SP (2006) Modelling and analysis of the atmospheric nitrogen deposition in North Carolina. Int J Glob Environ Issues 6:231–252
Rigge M, Smart A, Wylie B et al (2013) Linking phenology and biomass productivity in South Dakota mixed-grass prairie. Rangel Ecol Manag 66:579–587
Rigge M, Homer C, Wylie B et al (2019) Using remote sensing to quantify ecosystem site potential community structure and deviation in the Great Basin, United States. Ecol Indic 96:516–531
Riggs GA, Hall DK, Román MO et al (2015) MODIS snow products collection 6 user guide. Natl Snow Ice Data Cent, Boulder, CO, USA, p 66
Saraf S, John R, Boyte SP, Rigge MB (2023) Biophysical drivers for predicting the distribution and abundance of invasive yellow sweet clover in the Northern Great Plains. U.S. Geol Surv Data Release https://doi.org/10.5066/P9X08W4T
Shi T, Xu H (2019) Derivation of tasseled cap transformation coefficients for sentinel-2 MSI At-sensor reflectance data. IEEE J Sel Top Appl Earth Obs Remote Sens 12:4038–4048
Stoa TE (1933) Persistence of viability of sweet clover seed in a cultivated soil. J Am Soc Agron 25:177–181
Stohlgren TJ, Ma P, Kumar S et al (2010) Ensemble habitat mapping of invasive plant species. Risk Anal 30:224–235
Thornton MM, Shrestha R, Wei Y, Thornton PE, Kao SC, Wilson BE (2022) Daymet: daily surface weather data on a 1-km grid for North America, Version 4 R1. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/2129. Accessed on 01 May 2022
Thuiller W (2003) BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Glob Chang Biol 9:1353–1362
Thuiller W, Georges D, Engler R, Breiner F, Georges MD, Thuiller CW (2016) Package 'biomod2'. Species distribution modeling within an ensemble forecasting framework
Tikka PM, Koski PS, Kivelä RA, Kuitunen MT (2000) Can grassland plant communities be preserved on road and railway verges? Appl Veg Sci 3:25–32
Turkington R, Cavers PB (1979) The biology of Canadian weeds. 33. Medicago lupulina L. Can J Plant Sci 59(1):99–110
Van Riper LC, Larson DL (2009) Role of invasive Melilotus officinalis in two native plant communities. Plant Ecol 200:129–139. https://doi.org/10.1007/s11258-008-9438-6
Vermeire LT, Rinella MJ (2020) Fall water effects on growing season soil water content and plant productivity. Rangel Ecol Manag 73:252–258
Wolf J, Beatty S, Carey G (2003) Invasion by sweet clover (Melilotus) in Montane Grasslands, Rocky Mountain National Park. Ann Assoc Am Geogr 93:531–543
Wright CK, Wimberly MC (2013) Recent land use change in the Western Corn Belt threatens grasslands and wetlands. Proc Natl Acad Sci U S A 110:4134–4139
Wurtz TL, Macander MJ, Spellman BT (2010) Spread of Invasive Plants From Roads to River Systems in Alaska: A Network Model. U S For Serv Pacific Northwest Res Stn Gen Tech Rep PNW-GTR 699–708
Xiao X, Boles S, Frolking S et al (2002) Landscape-scale characterization of cropland in China using Vegetation and Landsat TM images. Int J Remote Sens 23:3579–3594
Zhao P, Lu D, Wang G et al (2016) Examining spectral reflectance saturation in landsat imagery and corresponding solutions to improve forest aboveground biomass estimation. Remote Sens. https://doi.org/10.3390/rs8060469