



# From Rangelands to Cropland, Land-Use Change and Its Impact on Soil Organic Carbon Variables in a Peruvian Andean Highlands: A Machine Learning Modeling Approach

Mariella Carbajal,<sup>1,2,3\*</sup>  David A. Ramírez,<sup>1\*</sup>  Cecilia Turin,<sup>4,5</sup>  
Sean M. Schaeffer,<sup>6</sup> Julie Konkel,<sup>7</sup> Johan Ninanya,<sup>1,8</sup> Javier Rinza,<sup>1</sup>  
Felipe De Mendiburu,<sup>9</sup> Percy Zorogastua,<sup>10</sup> Liliana Villaorduña,<sup>11</sup> and  
Roberto Quiroz<sup>12</sup>

<sup>1</sup>International Potato Center (CIP), Headquarters, P.O. Box 1558, 15024 Lima, Peru; <sup>2</sup>Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina, USA; <sup>3</sup>NC Plant Sciences Initiative, North Carolina State University, Raleigh, NC, USA; <sup>4</sup>Department of Animal Production, School of Animal Science, Universidad Nacional Agraria La Molina (UNALM), 15024 Lima, Peru; <sup>5</sup>Centro Experimental La Molina, Dirección de Supervisión y Monitoreo en Las Estaciones Experimentales Agrarias, Instituto Nacional de Innovación Agraria (INIA), Av. La Molina N° 1981, 15024 Lima, Peru; <sup>6</sup>Department of Biosystems Engineering and Soil Science, University of Tennessee, 2506 E.J. Chapman Drive, Knoxville, Tennessee 37996, USA; <sup>7</sup>Blount County Soil Conservation District, 1217 McArthur Road Maryville, Maryville, Tennessee 37804, USA; <sup>8</sup>Applied Meteorology Master Program, Universidad Nacional Agraria La Molina (UNALM), 15024 Lima, Peru; <sup>9</sup>Department of Statistics and Informatics, School of Economics and Planning, Universidad Nacional Agraria La Molina (UNALM), 15024 Lima, Peru; <sup>10</sup>Department of Agronomy, School of Agriculture, Universidad Nacional Agraria La Molina (UNALM), 15024 Lima, Peru; <sup>11</sup>Water Resources Engineering Master Program, Universidad Nacional Agraria La Molina (UNALM), 15024 Lima, Peru; <sup>12</sup>Sistema Nacional de Investigadores (SENACYT), Edificio 205, Ciudad Saber, Clayton, Panama

Received 17 April 2024; accepted 13 August 2024;  
published online 9 September 2024

**Supplementary Information:** The online version contains supplementary material available at <https://doi.org/10.1007/s10021-024-00928-7>.

**Author contributions** MC performed soil C analyses, ML modeling, SOC mapping, and led manuscript writing. DR led manuscript writing and the research team. CT led soil sampling, land use, and land cover categorization. SMS and JK established and led soil C analyses. JN and JR conducted statistical analysis and contributed to the writing. FDM, PZ, and LV designed and performed soil sampling. RQ led the research team, conceptualized the main study, and obtained financial support for the research. MC performed soil C analyses, ML modeling, SOC mapping, and led manuscript writing. DR led manuscript writing and the research team. CT led soil sampling, land use, and land cover categorization. SMS and JK established and led soil C analyses. JN and JR conducted statistical analysis and contributed to the writing. FDM, PZ, and LV designed and performed soil sampling. RQ led the research team, conceptualized the main study, and obtained financial support for the research.

\*Corresponding author; e-mail: mcarbaj@ncsu.edu; d.ramirez@cgjar.org

## ABSTRACT

Andean highland soils contain significant quantities of soil organic carbon (SOC); however, more efforts still need to be made to understand the processes behind the accumulation and persistence of SOC and its fractions. This study modeled SOC variables—SOC, refractory SOC (RSOC), and the <sup>13</sup>C isotope composition of SOC ( $\delta^{13}\text{C}_{\text{SOC}}$ )—using machine learning (ML) algorithms in the Central Andean Highlands of Peru, where grasslands and wetlands (“bofedales”) dominate the landscape surrounded by Junin National Reserve. A total of 198 soil samples (0.3 m depth) were collected to assess SOC variables. Four ML algorithms—random forest (RF), support vector machine (SVM), artifi-

cial neural networks (ANNs), and eXtreme gradient boosting (XGB)—were used to model SOC variables using remote sensing data, land-use and land-cover (LULC, nine categories), climate topography, and sampled physical–chemical soil variables. RF was the best algorithm for SOC and  $\delta^{13}\text{C}_{\text{SOC}}$  prediction, whereas ANN was the best to model RSOC. “Bofedales” showed 2–3 times greater SOC ( $11.2 \pm 1.60\%$ ) and RSOC ( $1.10 \pm 0.23\%$ ) and more depleted  $\delta^{13}\text{C}_{\text{SOC}}$  ( $-27.0 \pm 0.44\%$ ) than other LULC, which reflects high C persistent, turnover rates, and plant productivity. This highlights the importance of “bofedales” as SOC reservoirs. LULC and vegetation indices close to the near-infrared bands were the most critical environmental predictors to model C variables SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ . In contrast, climatic indices were more important environmental predictors for RSOC. This study’s outcomes suggest the potential of ML methods, with a particular emphasis on RF, for mapping SOC and its fractions in the Andean highlands.

**Key words:** Artificial neural networks; Bofedales;  $^{13}\text{C}$  isotope composition; Extreme gradient boosting; Grasslands; Random forest; Refractory C fraction; Support vector machine.

## HIGHLIGHTS

- ML algorithms consistently modeled SOC variables with high performance
- Free publicly available remote sensing data was useful for SOC variables prediction
- Bofedales and grasslands were the most important reservoirs of SOC and fractions

## INTRODUCTION

The High Andes, located between 5°S and 20°S and above 4000 m.a.s.l., are characterized by their rich agro biodiversity (Monge-Salazar and others 2022) and ecosystem services (Rolando and others 2017a). However, the high melting rate of their glaciers (Zemp and others 2019), high frequency and intensity of extreme events (heavy rainfalls, frosts, strong winds, droughts, among others; Poveda and others 2020), and changes in land use (mainly agricultural intensification and encroach-

ment; Rolando and others 2017a) make these areas especially vulnerable to climate change. Rising temperatures have led to an expansion of crops to higher elevations (Skarbø and VanderMolen 2016), promoting an increasing incidence of pests (Dangles and others 2008) and diseases.

Global warming drives crop encroachment on the Andes’s higher lands (Rolando and others 2017a), which causes a substantial land-use change and the reduction of soil organic carbon (SOC) pools. External market demand, environmental policy, and management of high Andean grasslands have led to regrettable examples of landscape degradation and transformation. In the Andean highlands of Junin-Peru, the so-called “boom” of maca (*Lepidium meyenii*), a “superfood” appreciated for its energizing nutritional power with high demand in the Asian market during 2011–2015 (Turin and others 2018), has transformed a landscape dominated by highland grassland cover to a prevalence of bare soil degraded by maca cultivation. This cultivation process involves burning and plowing the grassland with heavy machinery, releasing significant amounts of carbon ( $123\text{--}136\text{ t ha}^{-1}$ ; Rolando and others 2017b). Furthermore, other activities put to risk the conservation and functioning of high Andes wetlands named “bofedales,” which are crucial for water security in lowlands (MINAM 2015) and for conserving significant soil C stocks (Monge-Salazar and others 2022; Hribljan and others 2016) and biodiversity (Polk and others 2019; Maldonado 2014). These activities involve extracting compact blocks of vegetation with a thin layer of soil, which is then used as alternative energy for heating and cooking (Caro and others 2014) and the overgrazing caused by domestic livestock (Cochi Machaca and others 2018).

Andean soil contains high quantities of SOC, the carbon that remains in the soil after the partial decomposition of organic matter by microorganisms (Alavi-Murillo and others 2022). However, few studies address SOC assessment and modeling in the Andean highlands region. Refractory SOC (RSOC) represents a fraction that persists in soil and has a finite turnover time of thousands of years (Krull and others 2003). It represents one of the significant global SOC pools (Jagadamma and others 2010), and its quantification is crucial for understanding C dynamics (decomposition and stabilization processes). Also, the  $^{13}\text{C}$  isotope composition of SOC ( $\delta^{13}\text{C}_{\text{SOC}}$ ) constitutes another cru-

cial soil trait because it may be used to estimate plant inputs into soil organic matter (Ehleringer and others 2000; Bernoux and others 1998). Moreover,  $\delta^{13}\text{C}_{\text{SOC}}$  has been shown to vary with SOC turnover rate, and sources of SOC under land-use change (Ehleringer and others 2000; Xia and others 2021; Han and others 2023). Predictions of quantity and turnover rate based on  $\delta^{13}\text{C}_{\text{SOC}}$  are subject to errors associated with climate variability, temporal differences, and anthropogenic contamination. Therefore, it is essential to quantify these errors and compare them with other results to achieve robustness. Artificial intelligence methods, including machine learning (ML) and deep learning, emerged in the last two decades in pedometrics and have been demonstrated to outperform other SOC modeling approaches, such as linear regression and geostatistical approaches, due to their ability to find nonlinear patterns in a multidimensional set of potential environmental predictors (Somaratna and others 2016; Keskin and others 2019; Veronesi and Schillaci 2019; Chen and others 2022; Grunwald 2022; Zhu and others 2022). However, multiple algorithms are commonly tested and compared because no rule exists for choosing the best ML algorithm. This is because ML models are considered black boxes (the underlying processes for prediction are unknown), and the algorithms fit differently depending on the input data.

Remote sensing data have been used as a primary source of predictor variables. Multispectral imagery, including Landsat (Ayala Izurieta and others 2021), MODIS (Sreenivas and others 2016), SPOT (Liu and others 2015a), and others, is used as a nondestructive data source to study SOC variability (Gehl and Rice 2007; Chatterjee and others 2021) through the calculation of different spectral indices. The Andean highlands have received little attention for quantifying soil C fractions. No approaches for developing predictive models that help to understand C process dynamics and the main drivers for this system have been validated, perhaps due to the high spatial heterogeneity and limited resources to conduct sampling. In this study, SOC, RSOC, and  $\delta^{13}\text{C}_{\text{SOC}}$  (referred to as soil C target variables hereafter) were measured and used to develop predictive models using ML algorithms and publicly available remote sensing data in the Andean highlands of Junin-Peru. This study aims: i) to compare the soil C target variables among the most important land uses in the zone, ii) to analyze the performance of some ML methods for predictive modeling, iii) to find their most important environmental predictors related to land use, climate,

topography, and soil properties, and iv) to spatially model SOC across the study area.

## MATERIALS AND METHODS

### Study Area

The study was conducted in the central Peruvian Andean highlands within the districts of Junin and Carhuamayo in the department of Junin ( $10^{\circ} 01' \text{ S}$ ,  $76^{\circ} 07' \text{ W}$ , 4200 m a.s.l.). The study area comprised about 800 km<sup>2</sup> within the Junin National Reserve buffer zone (Figure 1), whose primary purpose is to protect the grassland and bofedal ecosystems and biodiversity of Junin's lake and the surrounding central Andean highlands. The Ramsar Convention identifies this site as an essential wetland area (site number 882; RSIS 2021). The climate is rainy and cold, with dry autumn/winter according to the Thornthwaite climatic classification system (SE-NAMHI 2022). The annual average maximum temperature, minimum temperature, and precipitation are 9–19 °C, -3–3 °C, and 500–1200 mm, respectively (period 1981–2010; SENAMHI 2022). The soil in the study area is characterized mainly by a predominance of Inceptisols with a trend of high SOC concentrations and acidic pH (Rolando and others 2018).

As grasslands and “bofedales” dominate the landscape, the primary land use and the main livelihood is grazing livestock consisting of cattle and sheep, which coexist with wild vicuñas (*Vicugna vicugna*). In some cases, subsistence agriculture is practiced with crops of potato and maca and is limited to a few small spots of land. However, from the 1990s to the present day, maca cultivation has had a significant expansion, becoming the primary driver of land-use change and the leading disruptor of the high Andean drylands (“puna”) ecosystem (Turin and others 2018).

### Measured Soil Data

The soil sampling sites were selected following the Latin Hypercube sampling (LHS) statistical method, which provides an efficient way of sampling variables, ensuring a good representation of the environmental characteristics of the study area (Carré and others 2007; Wang and others 2022; Stein 1987; McKay and others 1979). The LHS method used the multidimensional distributions of the slope, precipitation, minimum and maximum temperatures, normalized difference vegetation index (NDVI), and land cover estimated by a supervised classification from Landsat 8 imagery from the United States Geological Survey (USGS



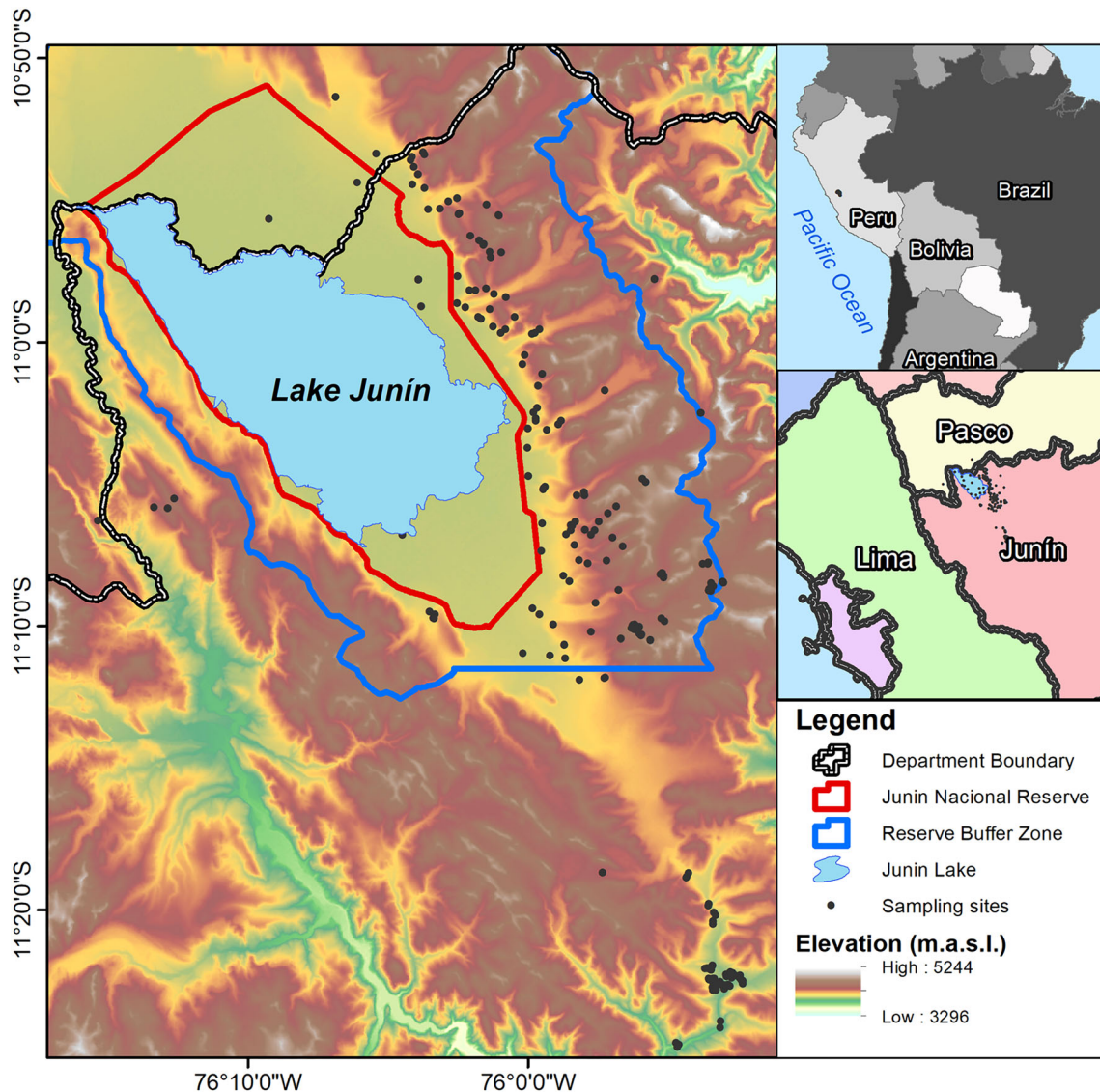


Figure 1. The study area and the 198 soil sample locations in the proximities of Junin Lake are located in the Central Peruvian Andean Region of the Province of Junin. Advanced Land Observing Satellite (ALOS) Phased Array type L-band Synthetic Aperture Radar (PALSAR) digital elevation model was used for mapping.

2020) to determine the sampling locations. The sampling locations were adjusted in practice due to high slopes and accessibility, resulting in the selection of 198 sites. A composite soil sample (~ 5 kg) was gathered at each sampling site from five locations: one central point and four points positioned 2 m apart in the N, S, E, and W cardinal directions. These samples were collected using an auger from the 0.3 m soil profile (Art's Manufacturing & Supply Inc., model Mud Augers, USA). In addition, a pit measuring  $0.8 \times 0.7 \times 0.5 \text{ m}^3$  was dug at the central point for bulk density measurements within the 0–0.3 m soil profile, using metal cylinders of 0.05 m in diameter. Then, C stock was

estimated by multiplying SOC (see its determination below) with bulk density following Rolando and others' (2017b) procedure. Unfortunately, bulk density measurements were made for just 64% of the sites selected due to operational inconveniences; therefore, LULC averaged values are reported.

Composite soil samples were analyzed for texture and pH using a hydrometer and suspension potentiometer (water in 1:1 relation) at the Soil Laboratory of the National Agrarian University La Molina—Lima, Peru. The soil C target variables' values were determined using a Combustion Module coupled to a Cavity Ring-Down Spec-

trospectroscopy (CM-CRDS) system based on Liu and others's (2018) procedure for SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ . Thus, a soil subsample per site was sieved to < 2 mm, dried at 60 °C, and ground with a mortar and pestle. Then, the final soil sample weights to be analyzed were determined by land-use type based on their mean expected soil C concentration. Hence, 0.015, 0.030, 0.027, and 0.0075 g were packaged in tin capsules for maca crops, fallow and cultivated pastures, native grasslands and improved pastures, and wetlands ("bofedales," see below), respectively. For RSOC, a second soil subsample per site was oxidized using  $\text{H}_2\text{O}_2$ , according to Jagadamma and others (2010), with slight modifications. Thus, 1 g of sieved soil (< 2 mm) was oxidized by adding 90 ml of 10%  $\text{H}_2\text{O}_2$  for 2–3 days, centrifugated for 15 min, washed three times with deionized water, and freeze-dried. From the remaining soil, 0.075 g was weighed and packaged in tin capsules. Finally, all tin capsules were submitted to a CM-CRDS system (G2131-iAnalyzer, Picarro Inc., USA).  $\delta^{13}\text{C}_{\text{SOC}}$  was estimated from the  $^{13}\text{C}/^{12}\text{C}$  natural abundance values reported by the equipment relative to international standard VPDB (Vienna Pee Dee Belemnite) using the equations by Liu and others (2018). All the analyses were performed in the Schaeffer Lab in the Department of Biosystems Engineering and Soil Science at the University of Tennessee, Knoxville, USA.

### Environmental Predictors and Land-Use and Land-Cover Categories

Given that the soil C target variables result from complex processes and interactions of several environmental factors—including topography, climate, soil properties, and vegetation—the primary environmental predictors underpinning their unique processes are likely to vary in significance. Despite this complexity and considering the limited ML experience in predicting soil C variables beyond SOC, this study utilized an identical set of features (environmental predictors hereafter) for SOC, RSOC, and  $\delta^{13}\text{C}_{\text{SOC}}$ . Thus, the environmental predictors considered for the models were obtained from publicly available remote sensing data, soil lab analysis, and vegetation type and condition at soil sampling (see definitions in Table 1). The topographic variables were elevation (DEM—Digital elevation model), slope, aspect, and topographic wetness index (TWI), derived from the Advanced Land Observing Satellite (ALOS) Phased Array type L-band Synthetic Aperture Radar (PALSAR)—Radiometric Terrain Correction prod-

uct. The climate indices were the minimum and maximum of the average monthly minimum (TMNN and TMNX, respectively) and maximum temperatures (TMXN and TMXX, respectively) and the average annual total precipitation (PREC), calculated from WorldClim version 2.1 climate data (period 1970–2000 with ~ 1 km resolution). Vegetation also plays a vital role in these carbon variables, so the nine spectral bands and several vegetation indexes were estimated from a Landsat 8 Operational Land Imager (OLI) imagery from November 26th, 2014 (see list in Table 1). Remote sensing data was preprocessed using Environmental Systems Research Institute (ESRI) ArcGIS software (ESRI, 2011, Redlands, CA).

In addition, as the predominant vegetation was grasslands and grasslands converted into maca fields, finer land-use and land-cover (LULC) categories were defined depending on the type of grassland, condition, and history (see Figure 2). "Vigorous grasslands" (n = 45) was defined as healthy, tall grasslands with good cover and sparse bare soil. "Partially degraded grasslands" (n = 57) were referred to as medium-sized, sparse grasslands with some bare soil, whereas "degraded grasslands" (n = 47) were typified as low and sparse grasslands covered surrounded by abundant bare soil. All grassland categories are land used neither for cropping activities nor perturbed. "Improved pastures" (n = 5) referred to grasslands with introduced cultivated species such as white clover (*Trifolium pratense*) and red clover (*Trifolium repens*) through inter-seeding, implying a minimum perturbation since it does not require plowing. "Cultivated pastures" (n = 24) was defined to transform native grasslands into an association of species such as king grasses (*Lolium multiflorum*, *Lolium perenne*) and clovers (*Trifolium spp.*), introduced 40 years ago in the case of the multi-communal cooperative system and 15 years ago in the farmer community system. "Bofedales" (n = 10) is a type of Andean highland wetland with hydromorphic vegetation and generally accumulates peat, seasonally or permanently saturated with water (Monge-Salazar and others 2022). "Fallow 1" (n = 13) referred to bare soils from recently harvested maca crops or up to 2 years of fallow, which in turn come from the recent conversion of vigorous or partially degraded grasslands plowed to be converted to maca cropland. "Fallow 2" (n = 20) was composed of bare soils with invasive sparse grass species, coming from maca crops harvested 3 to 5 years ago, which in turn result from the conversion of "vigorous" or "partially degraded grasslands" that have been plowed to be transformed into maca cropland.

**Table 1.** Features Considered as Potential Environmental Predictors in Soil Organic Carbon Prediction

Variables	Abbreviation	Equation	Source
Land-use and land-cover condition at sampling	LULC		Field sampling
Soil texture particle sizes: sand content (%)	SAND		
Soil texture particle sizes: silt content (%)	SILT		
Soil texture particle sizes: clay content (%)	CLAY		
pH in water	pH		
Closest Euclidean distance to the lake	DLAKE	$d(\text{lake, sample})$	
Average annual precipitation (mm)	PREC		Climate variables: WorldClim v2.1, climate data from 1970–2000. 30 s resolution (~ 1 km)
Minimum of average monthly minimum temperature (°C)	TMNN	$\min(\text{Tmin})$	
Maximum of average monthly minimum temperature (°C)	TMNX	$\max(\text{Tmin})$	
Minimum of average monthly maximum temperature (°C)	TMXN	$\min(\text{Tmax})$	
Maximum of average monthly maximum temperature (°C)	TMXX	$\max(\text{Tmax})$	
Global-Aridity index	ARID		CGIAR Consortium for Spatial Information (CSI)
Global-Potential evapotranspiration	PET		
Ultra-blue band (435 – 451 nm)	UBLUE		Landsat 8 Operational Land Imager (OLI)
Blue band (452 – 512 nm)	BLUE		
Green band (533 – 590 nm)	GREEN		
Red band (636 – 673 nm)	RED		
Near-infrared band (851 – 879 nm)	NIR		
short-wave infrared-1 band (1566 – 1651 nm)	SWIR1		
Short-wave infrared-2 band (2107 – 2294 nm)	SWIR2		
Spectral vegetation indexes 1	SER1	$\text{Red}/\text{Green}$	
Spectral vegetation indexes 2	SER2	$\text{Red}/\text{SWIR2}$	
Spectral vegetation indexes 3	SER3	$\text{SWIR1}/\text{SWIR2}$	
Normalized Difference Vegetation Index	NDVI	$(\text{NIR} - \text{Red})/(\text{NIR} + \text{Red})$	
Enhanced Vegetation Index	EVI	$2.5 \times \left( \frac{\text{NIR} - \text{Red}}{\text{NIR} + 6 \times \text{Red} - 7.5 \times \text{Blue} + 1} \right)$	
Soil-Adjusted Vegetation Index. L = 0.5	SAVI	$(1 + L) \times ((\text{NIR} - \text{Red})/(\text{NIR} + \text{Red} + L))$	
Modified Soil Adjusted Vegetation Index	MSAVI	$\frac{2 \times \text{NIR} + 1 - \sqrt{(2 \times \text{NIR} + 1)^2 - 8 \times (\text{NIR} - \text{Red})}}{2}$	
Normalized Difference Moisture Index	NDMI	$(\text{NIR} - \text{SWIR1})/(\text{NIR} + \text{SWIR1})$	
Normalized Burn Ratio 1 and 2	NBR1	$(\text{NIR} - \text{SWIR2})/(\text{NIR} + \text{SWIR2})$	ALOS PALSAR* Derived topographic properties from DEM
Normalized Burn Ratio 1 and 2	NBR2	$(\text{SWIR1} - \text{SWIR2})/(\text{SWIR1} + \text{SWIR2})$	
Digital Elevation Model	DEM		
Slope in degrees	SLOPE		
Aspect	ASPECT		
Topographic Wetness Index**	TWI	$\ln(a/\tan\beta)$	

(\*) Advanced Land Observing Satellite (ALOS) Phased Array type L-band Synthetic Aperture Radar (PALSAR).

(\*\*) Calculated according to Quinn and others (1991), where  $a = (\text{Total catchment area})/(\text{Flow width})$  and  $\beta = \text{Slope}$ .



“Fallow-3” ( $n = 12$ ) referred to invasive grass species with sparse low vegetation resulting from long-standing maca fallow ( $> 5$  years) of transformed grasslands into maca cropland. (Table 2).

## Modeling Approach

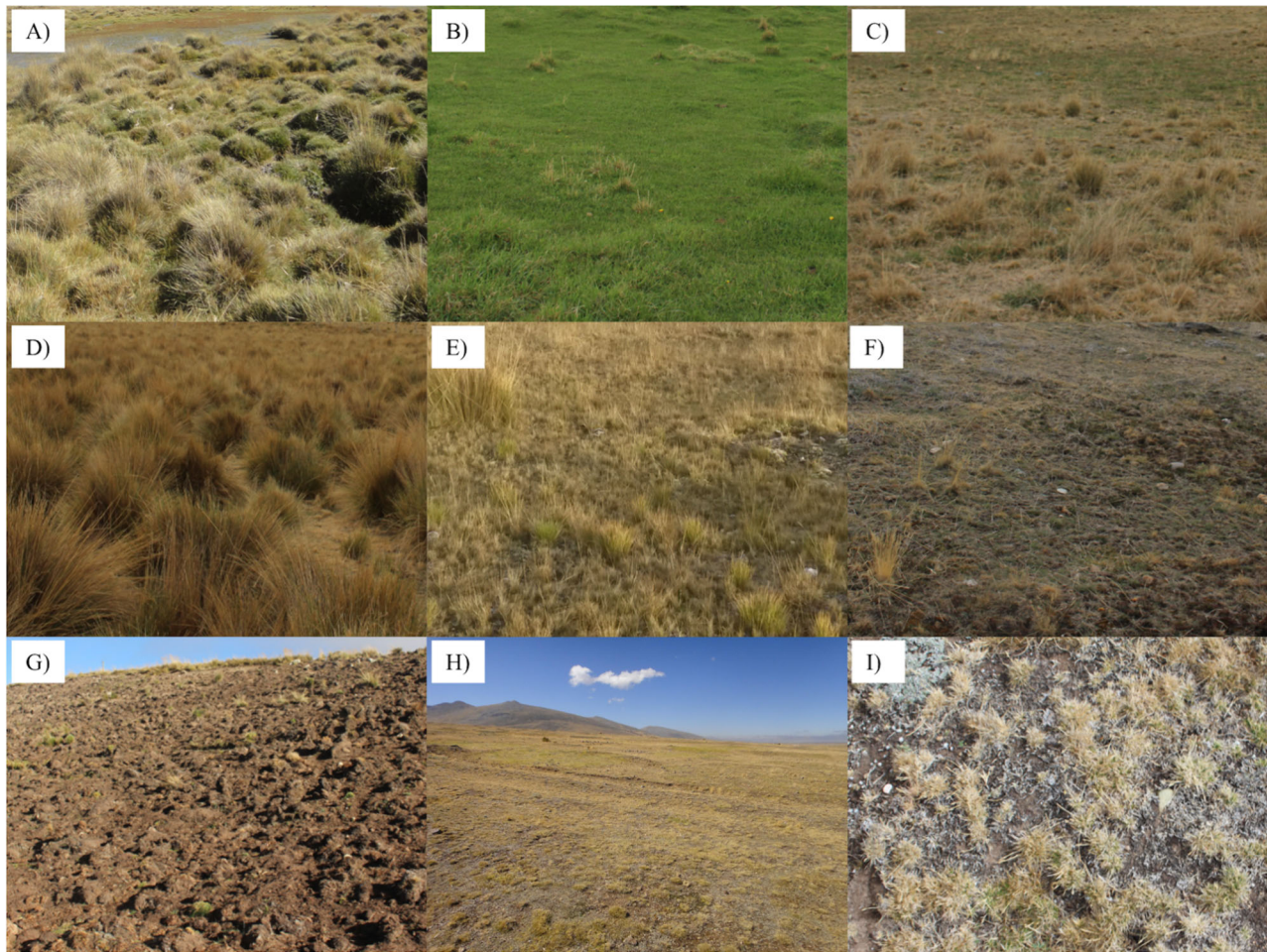
From the 42 potential environmental predictors considered for this study (33 numerical and nine categorical from LULC, Table 1), some may be nonessential or repetitive, and it is always better to identify and exclude them from the model building. Addressing and preselecting the minimum-optimal and all-relevant features to include (feature selection) helps optimize the model prediction and reduce overfitting (Parsaie and others 2021). Among the feature selection methods, the Boruta method (Kursa and Rudnicki 2010) has yielded better results when working with environmental processes like SOC decomposition due to its ability to identify linear and nonlinear relationships from complex processes (Keskin and others 2019; Zeraatpisheh and others 2022). This study used Boruta to select all-relevant and tentative environmental predictors for building the models for every soil C target variable. Based on a random forest (RF) classification algorithm, this method creates randomness in the system and determines the unimportant, meaningful, and tentative attributes of a given variable. After the Boruta feature selection, the new dataset underwent balancing and partitioning, including the selected environmental predictors and the soil C target variables. This partitioning for model training and testing was based on the values of the soil C target variable, utilizing a fivefold approach (Yates and others 2022). The approach comprises five cycles of model training and testing, where each iteration involves permuting four folds for training (75%) and reserving onefold for testing (25%). The four top-performing algorithms in predicting SOC (John and others 2020; Emadi and others 2020)—RF, artificial neural Networks (ANN), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB)—were employed to develop predictive models for the soil C target variables. Due to the differences in the ranges and distributions of the environmental predictors' values, feature scaling (transformations of values) through scaling (subtracting feature mean and dividing by feature standard deviation, mean 0, and standard deviation 1) and normalization (dividing by the feature maximum, range from 0 to 1), was executed and tested to determine the most effective method for enhancing model performance. Following the lit-

erature recommendation, especially for regression and when variable importance is of interest, feature scaling was applied even for the tree-based algorithms RF and XGB (Strobl and others 2007; Balabaeva and Kovalchuk, 2019). Then, hyperparameters were tuned using “out-of-bag,” “tenfold cross-validation repeated three times, and “leave-one-out cross-validation” resampling methods for RF, SVM, and ANN-XGB. For every soil C target variable modeled, performance metrics were averaged across the fivefold partitions for both the training and testing phases (see next section) and compared to identify the best predictive ML model. Once the best model was found and due to the small dataset, the ML model was retrained using the whole dataset (without partitioning), and the important variables were evaluated. The ML models were built and assessed using R 3.6.1 and the packages “Boruta” v7.0.0 (Kursa and Rudnicki 2010) for feature selection and “caret” v6.0.86 (Kuhn and others 2019) for applying the RF, ANN, SVM, and XGB algorithms.

Finally, its spatial distribution was mapped, and SOC was identified as the primary variable of interest. The RF model was recalibrated by retraining it, using the most important spatially available environmental predictors, which included LULC, SER2, NDMI, MSAVI, NDVI, DLAKE, SWIR2, and NBR1. For LULC, a land-cover classification was performed using the RF classification algorithm in Google Collaboratory. This classification used the 198 sample sites across the nine LULC categories and categories for water bodies, inundated areas, urban areas, rocks, and cattails (Mantas and Caro 2023) and the same Landsat imagery used in this study. These additional land-use categories were masked together and defined as “Non-carbon storing surfaces” for mapping purposes. Furthermore, a raster depicting the Euclidean distance—the shortest distance—to Junin Lake (DLAKE) was generated based on the lake's boundary. The remaining predictors were Landsat-based indices, which were already spatially available. The training samples for classification, the DLAKE raster, and the process of raster snapping (at 30 m resolution) for all variables were conducted in ArcGIS.

## Statistical Comparison of Soil Organic C Variables and Models Performance Assessment

The Kruskal–Wallis rank sum test was used to test significant differences among LULC for the soil C target variables, followed by Dunn's post hoc test



**Figure 2.** Photos of land-use and land-cover categories (see definition in Materials and Methods section): **A** Bofedales, **B** Cultivated pastures, **C** Improved pastures, **D** Vigorous grasslands, **E** Partially degraded grasslands, **F** Degraded Grasslands, **G** Fallow areas fallow with 0–2 years after maca cultivation (Fallow 1), **H** Fallow areas fallow with 3–5 years after maca cultivation (Fallow 2), **I** Fallow areas fallow with > 5 years after maca cultivation (Fallow-3).

with Holm’s correction method for adjusting *p*-values for multiple comparisons. For that analysis, the R packages “stats” (R Core Team 2022) and “DescTools” (Signorel and others 2022) were used. Next, the coefficient of determination ( $R^2$ ) and root mean square error (RMSE) were used to assess the performance of the ML models tested.  $R^2$  represents the proportion of variance explained by each ML model, and RMSE indicates the accuracy of the predicted values (Yang and others 2014).  $R^2$  and RMSE were calculated as follows:

$$R^2 = \left( \frac{N \sum_{i=1}^N X_i \hat{X}_i - \sum_{i=1}^N X_i \sum_{i=1}^N \hat{X}_i}{\sqrt{[N \sum_{i=1}^N X_i^2 - (\sum_{i=1}^N X_i)^2][N \sum_{i=1}^N \hat{X}_i^2 - (\sum_{i=1}^N \hat{X}_i)^2]}} \right)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\hat{X}_i - X_i)^2}{N}}$$

where  $\hat{X}_i$ ,  $X_i$ , and *N* are the model predicted values, observed values, and, total number of observed values, respectively. Higher  $R^2$  (close to 1) and lower RMSE (close to 0) mean better ML model performance. Model performance metrics were calculated as the average across the fivefold partitions for training and testing.

## RESULTS

### Soil C Measurements by LULC

SOC values ranged between 1.67–17.77%, with the lowest value found in “degraded grasslands,” which was significantly lower than that of “bofedales” (*p*-value < 0.001) and “partially degraded grasslands” (*p*-value < 0.01) (Figure 3A, Table S1). The highest SOC value was found in “bofedales,” being significantly (*p*-value < 0.05)



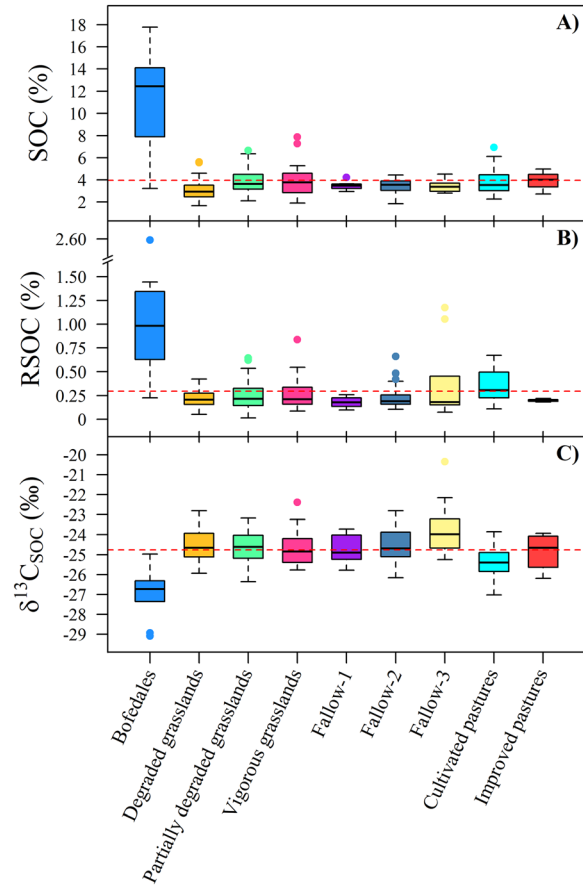
**Table 2.** Model Performance Metrics ( $R^2$ —Coefficient of Determination and RMSE—Root mean Square Error) for Random Forest (RF), Artificial Neural Networks (ANNs), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB) Algorithms on the Models' Training and Testing of Soil Organic Carbon (SOC), Refractory SOC (RSOC), and  $^{13}\text{C}$  Isotopic Composition of SOC ( $\delta^{13}\text{C}_{\text{SOC}}$ )

Soil C variable	Model	Training		Testing	
		RMSE	$R^2$	RMSE	$R^2$
SOC	RF	0.77	0.87	1.47	0.49
	ANN	1.15	0.70	1.49	0.43
	SVM	1.11	0.72	1.66	0.36
	XGB	0.92	0.81	1.53	0.42
$\delta^{13}\text{C}_{\text{SOC}}$	RF	0.35	0.89	0.79	0.42
	ANN	0.60	0.66	0.85	0.29
	SVM	0.42	0.84	0.88	0.26
	XGB	0.30	0.84	0.81	0.37
RSOC	RF	0.10	0.87	0.19	0.46
	ANN	0.12	0.80	0.18	0.50
	SVM	0.14	0.75	0.20	0.46
	XGB	0.05	0.95	0.20	0.41

2–3 times higher than that of the other LULC categories except for “improved pastures” (Figure 3A, Table S1). RSOC values ranged between 0.01 and 2.58%, being the lowest and highest ones found in “fallow-1” and “bofedales,” respectively (Figure 3B). “Cultivated” and “improved pastures” were not significantly ( $p$ -value > 0.05) lower than “bofedales” which was 2–3 times higher than the other LULC categories (Table S1). Values of  $\delta^{13}\text{C}_{\text{SOC}}$  ranged between  $-29.09$ – $20.35$  ‰ being the highest one found in “fallow-3” (Figure 3C). The lowest value was found in “bofedales” (all its values were below the overall mean of  $-24.76 \pm 0.074$  ‰) which was significantly different to all the other LULC categories except to “cultivated” and “improved pastures” (Figure 3C, Table S1). “Fallow-3” ( $p$ -value < 0.01), “degraded grasslands” ( $p$ -value < 0.01), and “partially degraded grasslands” ( $p$ -value < 0.05) showed significant differences compared to “cultivated pastures” (Table S1). Bulk density in “bofedales” was approximately half compared to other LULC categories ( $0.49 \text{ t m}^{-3}$  vs.  $0.98$ – $1.09 \text{ t m}^{-3}$ ), while carbon stock was nearly twice as high ( $210.9 \text{ t ha}^{-1}$  vs.  $97.4$ – $126.3 \text{ t ha}^{-1}$ ) (Table S2).

### Model Performance and Comparison

Overall, RF consistently outperformed other ML algorithms in modeling soil C target variables



**Figure 3.** **A** Soil organic carbon (SOC), **B** refractory SOC (RSOC), and **C**  $^{13}\text{C}$  isotopic composition of SOC ( $\delta^{13}\text{C}_{\text{SOC}}$ ) on different land-use and land-cover categories (see definition in Materials and Methods section). Red dashed horizontal line represents the global average.

models, achieving  $R^2 > 0.87$  during training, except for XGB in the RSOC model (0.95), and  $R^2 > 0.42$  during testing, except for ANN in the RSOC model ( $R^2 = 0.50$ ) (Table 2). Thus, following the criteria indicated in Sect. “Statistical Comparison of Soil Organic C Variables and Models Performance Assessment” and analyzing the average fivefold  $R^2$  and RMSE values for training and testing, respectively (Table 2), RF was selected as the most appropriate model for predicting SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ , and ANN for predicting RSOC.

### Explanatory Variables

The environmental predictors excluded (see selection criteria in Sect. “Modeling Approach”) from the model building of the soil C target variables were SILT (silt content), CLAY (clay content), BLUE (blue band), GREEN (green band), SWIR1 (short-wave infrared-1 band), SLOPE, ASPECT,

TWI, TMNN, TMNX, and “Cultivated pastures” (data not shown). A total of 23, 22, and 20 out of the 42 environmental predictors were selected (see selection criteria in Sect. “Modeling Approach”) for building the models for SOC,  $\delta^{13}\text{C}_{\text{SOC}}$ , and RSOC, respectively (data not shown). From the selected environmental predictors, “bofedales” were identified as the most critical for SOC, followed by SER2 (Spectral vegetation indexes 2) and NDMI (Normalized Difference Moisture Index), both of which were considerably less important (Figure 4A). Regarding  $\delta^{13}\text{C}_{\text{SOC}}$ , “bofedales” also were the most critical environmental predictor, followed by NDMI and DLAKE, which had similar importance, and then by NIR (Near-infrared band) and pH, which were the next ones in importance (Figure 4B). For RSOC, pH was as critical as “Fallow-3,” followed by SWIR2 (short-wave infrared-2 band), “bofedales,” and EVI (Enhanced Vegetation Index) with lower importance (Figure 4C).

### SOC Mapping

The land-cover classification yielded accuracies of 95% during training and 60% during testing. Predicted SOC values within the study area ranged from 2.7 to 11.5% (Figure 5). The highest SOC values were predominantly found north and south of Junin’s lake, mainly in the “bofedales” zone. Areas with the next highest SOC values were found

in the southernmost part of the study area, primarily corresponding to cultivated pasture zones. Conversely, the lowest SOC values were predicted in the western Reserve Buffer Zone.

## DISCUSSION

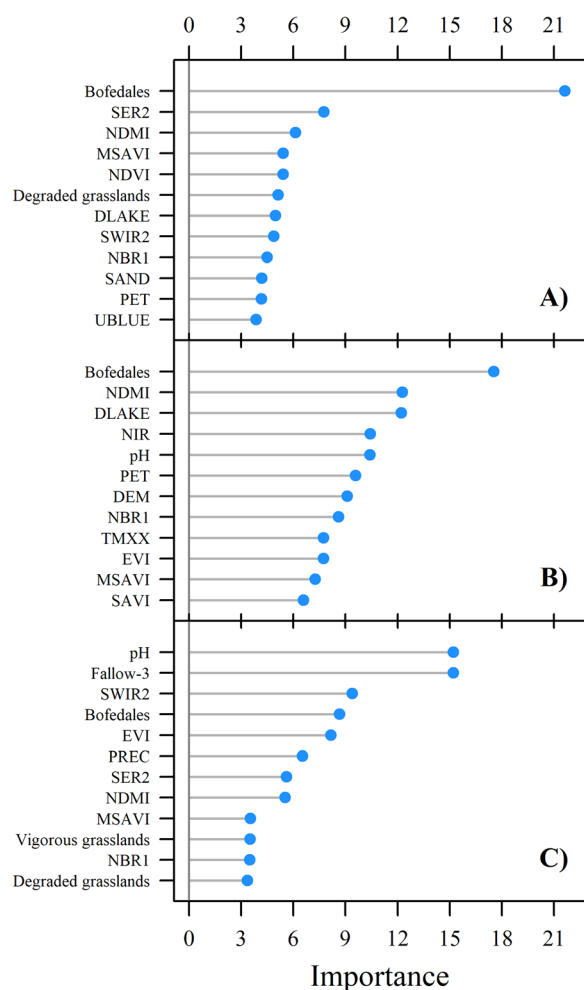
### “Bofedales” as Essential Reservoirs of Soil Organic Carbon in the Andes Highlands

Bofedales showed higher SOC amounts compared to other assessed land uses (Figure 2). Even though these wetlands have been recognized as an essential reservoir of SOC in the Andes (Alavi-Murillo and others 2022; Segnini and others 2013), their relevance in policy incidences and conservation/restoration actions is scarce, or null (Maldonado 2014). The SOC range of values in this study (3.2–17.8%) was in the lower range of values reported by other studies (13.2–83.2%) (Cooper and others 2010; Segnini and others 2010; Alavi-Murillo and others 2022; Monge-Salazar and others 2022). Plant biomass extraction from the soil through “champeo” and overgrazing has been reported in the study area (Caro and others 2007, 2014; Salvador and others 2014; Mantas and Caro 2023); these perturbations could promote SOC reduction. On the other hand, C stock values (in the 0–0.3 m soil profile) found in this study (Table S2) were

**Table 3.** Caused Impacts by the Primary Land-Use Changes in the Study Area and Highland Andean Ecosystems Reported by Literature

The main land-use changes	Drivers in the study area	Effect	
		Soil/land components & function	Ecosystem services
• Native “bofedales”/grasslands to Degraded “bofedales”/grassland	Fuel demand and inappropriate management (“champeo”, overgrazing)	<b>SOM reduction</b> <sup>a</sup> , erosion <sup>b</sup> , loss of plant diversity and soil degradation <sup>c,d</sup> , water retention <sup>d</sup> , <b>loss of productivity</b> <sup>e</sup>	↓ FWFP, ↓ CS, ↓ NC
• Native grasslands to cultivated grassland	Agricultural policy reforms	> soil aggregation and <b>potential soil C sequestration</b> <sup>f</sup> , > <b>plant productivity</b> and soil fertility <sup>f,g</sup>	↑ FWFP, ↑ CS, ↑ SF, ↑ NC
• Native grasslands to maca crop	External market demand	Soils macroaggregate disruption <sup>f</sup> , > nutrient mineralization and soil fertility <sup>g</sup>	↑ FP, ↑ SF
• Maca crop to the fallow areas	External market demand	<b>SOC reduction</b> <sup>f</sup> , slow plant recovery <sup>f</sup> , > proneness to erosion and runoff <sup>f,g</sup>	↓ FWFP, ↓ CS, ↓ NC, ↓ WR

<sup>a</sup>Adler and Morales (1999), <sup>b</sup>Rolando and others (2017a), <sup>c</sup>Catorci and others (2014), <sup>d</sup>Cochi Machaca and others (2018), <sup>e</sup>Caro and others (2014), <sup>f</sup>Rolando and others (2017b), <sup>g</sup>Rolando and others (2018). Likely positive (↑) and adverse (↓) effects on provisioning (FWFP = food, wool, and fiber provision, FP = food provision), regulating (CS = carbon sequestration, WR = water regulation), and supporting (SF = soil fertility, NC = nutrient cycling.) ecosystem services are referred based on Millennium Ecosystem Assessment (Corvalán and others 2005). In bold are the processes directly and indirectly observed/measured in this study.



**Figure 4.** Rankings of the top most important environmental predictors defined for the best-performed machine learning model for soil organic carbon (SOC) with random forest (**A**, in %),  $^{13}C$  isotopic composition of SOC ( $\delta^{13}C_{SOC}$ ) with random forest (**B**, in ‰), and refractory SOC (RSOC) with Artificial Neural Network (**C**, in %). Importance is defined as the increase in the MSE prediction when the variables are permuted. The environmental predictors are described in Table 1.

slightly lower than those values reported in the literature for “bofedales” (211 vs. 230–306 t ha<sup>-1</sup> from Segnini and others 2010), grasslands (102–126 vs. 135–144 t ha<sup>-1</sup> from Farley and others 2013), fallows (106–11 vs. ~ 123 t ha<sup>-1</sup> from Rolando and others 2017b) and pastures (97–119 vs. 136 t ha<sup>-1</sup> from Rolando and others 2017b). There were no significant differences in C stock values among LULC categories except for “bofedales”, which was almost twice as high (Table S2). In “bofedales,” C stocks are more extensive and profound than the other LULC categories and range from 30–700 t C ha<sup>-1</sup> per meter of peat depth (peat

thickness can reach as deep as 15 m). The study findings highlighted the importance of “bofedales” as a reservoir of SOC and its stable C fractions and called for its conservation and restoration (see [Scale, Reach, and Impacts of Land-use Changes and their Implications for Conservation](#) Section).

The highest depletion of  $\delta^{13}C_{SOC}$  (ranged from – 29.5 to – 25.0 ‰) in “bofedales” than other LULC suggested that SOC was formed from plants under no water restriction conditions and better photosynthetic performance discriminating against  $^{13}C$  (Farquhar and others 1989; More and others 2022). This finding highlights the potential for relatively high primary productivity in “bofedales” in this Andean ecosystem. On the other hand, high enrichment of  $\delta^{13}C$  is also related to higher fractions of persistent SOC pools (Ehleringer and others 2000), which is consistent with our findings considering that “bofedales” showed the highest RSOC (1.10 ± 0.23%) than other LULC (Figure 3C). Furthermore, Segnini and others (2010) found an increase in persistent SOC pools with soil depth in Andean- “bofedales.”

Highland grasslands have been reported as other important reservoirs of C stocks and SOC in the Andes (Gibbon and others 2010; Zimmermann and others 2010; Farley and others 2013). In our study area, Rolando and others (2017b) detected that cultivated pastures showed similar values of SOC but a higher depletion of  $\delta^{13}C$  (4.5 ± 0.2‰ and – 26.0 ± 0.1 ‰, respectively) than native grasslands (4.6 ± 0.3‰ and – 25.6 ± 0.1 ‰, respectively) and fallow areas (4.1 ± 0.3‰ and – 25.6 ± 0.1 ‰, respectively). This has been interpreted as a higher depletion of  $\delta^{13}C$  in cultivated pastures from incorporating N-fixer species (white clover) and long-standing perennial grasses (like ryegrass), manure, and supplemental irrigation. In this study (in agreement with Rolando and others 2017b), “cultivated pastures” LULC showed significantly more depleted  $\delta^{13}C_{SOC}$  (– 25.4 ± 0.16 ‰) than “partially degraded” (– 24.6 ± 0.12 ‰) and “degraded grasslands” (– 24.5 ± 0.12 ‰), and fallows area after three years (– 23.7 ± 0.46 ‰) (Figure 3). This result suggested that vegetation that formed SOC in cultivated pastures had better physiological performance and that soil in degraded grasslands and fallow areas likely had more labile C forms.

### RF as Promising ML Algorithm for Predicting Soil C Variables in the Andean Highlands

Overall, among the ML algorithms, RF performed the best, capturing C processes’ nonlinear interac-



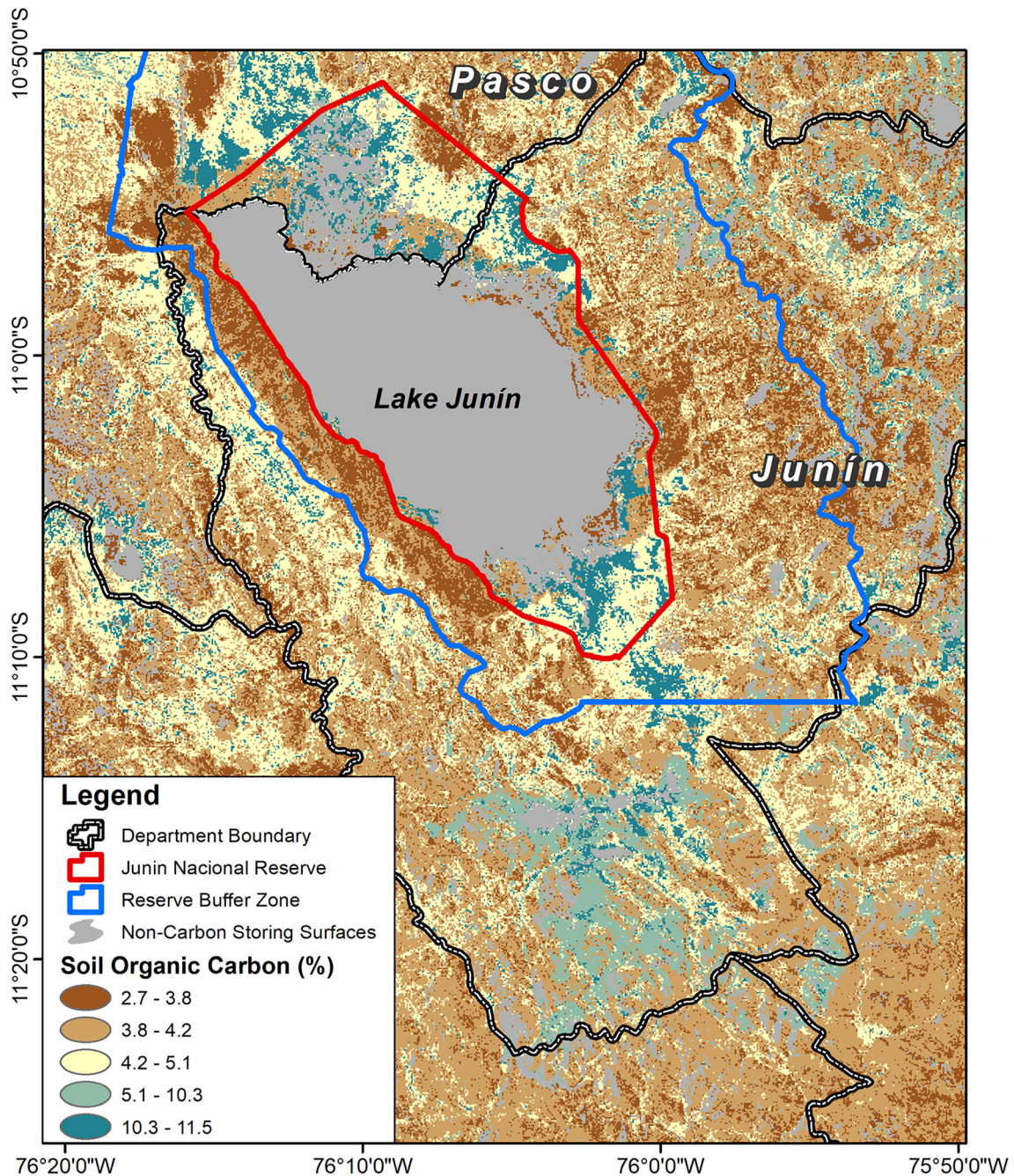


Figure 5. Predicted spatial distribution of soil organic carbon across the Lake Junín Region, Junín, Peru. This map showcases the distribution, as inferred by a random forest algorithm utilizing the eight most significant environmental predictors available spatially. Non-carbon storing surfaces correspond to classes such as water bodies, inundated surfaces, cattails, rocks, and urban areas.

tions with acceptable and consistent  $R^2$  and RMSE performances (Table 2), which agrees with most of the reported SOC modeling studies. In the literature, the performance of ML algorithms predicting SOC is highly variable. It depends on multiple factors, like the observed sample size, number and type of covariates, time–space resolution, extent of

the study area, and model algorithm (Grunwald 2022). Sample size has a more significant effect than the model algorithm on the model performance (Somaratna and others 2017).  $R^2$  is among the most reported model performance indicators for ML regression algorithms for soil C models due to its more straightforward interpretation, especially

when comparing multiple site applications where target value ranges and/or units may differ to use RMSE. However, most of these  $R^2$  values ranged from 0.24 to 0.68 (from first to third quartile) (Grunwald 2022), reflecting little understanding of the main drivers and methods for predicting SOC. For this study, the  $R^2$  of predicted soil C target variables varied from 0.42 to 0.50 for the best ML algorithms, agreeing with other studies with small sampling sizes and similar covariates (Zeraatpisheh and others 2022). Using multi-temporal data or soil nutrient indicators as covariates has been a strategy to counter the effect of a small sample size, allowing somewhat higher  $R^2$ , 0.58–0.68 (John and others 2020; Shafizadeh-Moghadam and others 2022). Therefore, the moderate performance of the models, especially in predicting  $\delta^{13}\text{C}_{\text{SOC}}$ , suggested that the processes involved are too complex for the given small sample size and/or some essential variables at the correct time–space scale were missing as covariates. Regarding the RMSE, predicting SOC got 1.47%, which seems high, but considering the small sample size and high SOC values from “bofedales,” it is fair and in the mid-range of the reported values from 0.59 to 2.7 across multiple SOC studies (Padarian and others 2019; Peng and others 2015; Safanelli and others 2020). Few studies modeled other C fractions apart from SOC with ML techniques; for example, Adi and Grunwald (2020) and Keskin and others (2019) modeled persistent C fraction at 0–0.2 m depth for Florida State using 850 and 1014 soil samples and 151 and 327 environmental predictors, respectively. When employing the RF algorithm, these studies achieved acceptable  $R^2$  values of 0.68 and 0.72, respectively. This suggests that model performances could be improved by adding sampled data and potential environmental predictors. The ANN model was selected for RSOC predictions due to its balanced performance in the training and testing phases. Although XGB and RF demonstrated superior learning capabilities during training, ANN performed well in training and exhibited the best generalization to unseen data in the testing phase (Table 2).

### Vegetation and Climatic Indices as Essential Predictors of Soil Organic Carbon

Quality and quantity of SOC are mainly determined by a soil’s physical and chemical environment, physical accessibility of organic matter to biological agents (that is, microbes and/or enzymes), and the ratio of C inputs to losses (Krull

and others 2003; Luo and others 2017; Sing and others 2018; Dynarski and others 2020). Even though land use significantly affects both labile and persistent C pools (Liu and others 2020; Padbhushan and others 2022; Smith 2008), the latter responds much slower than labile C pools to land-use and other human-induced changes (for example, land management) (Dynarski and others 2020; Padbhushan and others 2022; Sainepo and others 2018). Thus, LULC was one of the leading environmental predictors, “bofedales” the most relevant for SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ , and “Fallow-3” for RSOC. Several studies have highlighted the importance of LULC as a predictor variable for SOC (Emadi and others 2020; Keskin and others 2019; Xiong and others 2014) and RSOC (Keskin and others 2019; Xiao and others 2022) using ML algorithms. Regarding  $\delta^{13}\text{C}_{\text{SOC}}$ , Wang and others (2015) stress that the litter quality and soil water can increase the carbon isotope fractionation during organic matter decomposition. Because soil  $^{13}\text{C}$  isotope composition ( $\delta^{13}\text{C}$ ) is strongly influenced by leaf (litter)  $\delta^{13}\text{C}$ , variations in this variable can be influenced by LULC because it determines the type and quality of litter inputs into the soil (Smith and Chalk 2021; Wang and others 2013). Thus,  $\delta^{13}\text{C}$  values in labile C pools (that is, relatively “new” material) would reflect  $\delta^{13}\text{C}$  values closer to the current vegetation, whereas  $\delta^{13}\text{C}$  values in persistent C pools (that is, older material) shows relatively enriched  $\delta^{13}\text{C}$  values due to isotopic discrimination of the heavy isotope in soil organic matter compounds (Wang and others 2013). In addition, the crucial role of soil water and soil temperature and pH during soil organic matter decomposition has been highlighted as they increase the activity of soil fauna and microorganisms (Wang and others 2013; Wang and others 2015; Smith and Chalk 2021). Thus, we found that for “bofedales,” some indicators of soil water (DLAKE and NDMI) and vegetation (SER2), and pH were relevant environmental predictors for  $\delta^{13}\text{C}_{\text{SOC}}$  and RSOC (Figure 3). The greater relevance of pH for RSOC and  $\delta^{13}\text{C}_{\text{SOC}}$  than for SOC could be due to its impact on the activity and growth of microorganisms, which metabolize the different forms of C, resulting in a variation in the organic carbon isotopic composition of the soil (Neina 2019; Klink and others 2022). Also, soil pH can affect the interactions between soil minerals and organic matter, which determines the preservation and stability of C (Neina 2019). Although some soil variables, such as clay content, were reported as essential predictors for SOC (John and others 2020; Davy and Koen 2013), in this study, it was not of



high relevance, likely due to the importance of pH against other chemical indicators to explain SOC in Andean highlands soils (Alavi-Murillo and others 2022).

The relationship between SOC and remotely sensed and easily accessible variables has rarely been reported (Mirchooli and others 2020). However, Lamichhane and others (2019) reported that these variables were among the top five for SOC prediction. NDMI is a vegetation index that detects vegetation water content and is a good predictor for measuring SOC using ML methods (John and others 2020). Mirchooli and others (2020) found that coloration index and NDMI are the most critical environmental predictors for SOC prediction in the RF model, followed by elevation, NDVI, and slope. NDMI is indirectly related to soil moisture in the surface layers (0–0.3 m), and the latter can prevent the net loss of organic soils through oxidation (Liu and others 2015b). In this study, NDMI was the main environmental predictor in both SOC and  $\delta^{13}\text{C}_{\text{SOC}}$  under the RF model, followed by SER2, NIR, and NBR1. These last variables are closely related by the NIR and SWIR2 bands, found in the spectrum's wavelengths from 850 to 2200 nm. Bishop and others (2008) found a strong absorption near 1400 nm (also for Kaolinite) and 1900 nm, indicating the presence of water bound in the interlayer lattices of soil. This could provide the conditions for a physical protection mechanism through the interaction of SOC with the soil mineral matrix and the stabilization process by aggregate formation (Krull and others 2003). Also, Alabbas and others (1972) reported an inverse relationship to SOC approximately near this region of the spectrum, and with all this, it could have obtained the affinity to be one of the best environmental predictors for SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ .

### Scale, Reach, and Impacts of Land-use Changes and Their Implications for Conservation

The extraction of vegetation and part of the topsoil of “bofedales” and grasslands (an activity locally called “champeo”) has been carried out for decades by rural inhabitants (Caro and others 2014). “Champeo” allows the local population to guarantee fuel for domestic use (mainly cooking); however, it also constitutes a critical perturbation affecting SOC accumulation (Table 3). Overgrazing caused by domestic livestock is another activity reported in the study area (Caro and others 2007; Salvador and others 2014) that reduces peat production and can affect SOC pools from the assessed

“bofedales” and grasslands. Both perturbations (“champeo” and overgrazing) are the most important drivers that impact the change from vigorous/native to degraded “bofedales”/grasslands, reducing SOC (Figure 3) and provisioning, regulation, and supporting ecosystem services (Table 3). Land policy reforms during the '70s promoted establishing a multi-communal agrarian company (SAIS Tupac Amaru) in the region, covering more than 0.2 Mha, to increase grassland productivity for livestock (Diez 2020). Through these reforms, the natural grasslands from these lands were managed by incorporating productive pastures (ryegrass-white clover), irrigation, inorganic–organic fertilization, and rotational livestock grazing (Rolando and others 2017b). The land-use change from native to cultivated grasslands was the only one that was not considered a perturbation; it increased plant productivity (see first Discussion section) and soil health, promoting provisioning, regulation, and supporting ecosystem services (Table 3). Land-use changes caused by crop encroachment in highland grasslands are considered one of the most critical perturbations that threaten the ecosystem services of these landscapes in the highland Andes region (Tovar and others 2013; Rolando and others 2017a). Climate change facilitating the upward expansion of agriculture (Tovar and others 2012; Arce and others 2019) and socioeconomic factors like the increase of international market demand (like quinoa, Gamboa and others 2020) have been crucial drivers of Andes grasslands transformation.

In the study area, maca (*Lepidium meyenii*) cultivation was gradually extended in the grasslands of Junin since the early 90s for local, American, and European markets. Still, its expansion was massive in 2011–2015 to cover the high demands of the Asian markets. This led to a rapid transformation of the high Andean landscape with direct consequences on “puna” ecosystem services, such as the decrease of grassland primary production, reduced grazing areas, reduced land cover, loss of water infiltration and retention capacity of soils, besides changes in the main livelihood (Turin and others 2018) (see Table 3). This study corroborates findings previously reported in the field (Rolando and others 2017b, 2018), highlighting the occurrence of a degradation process following maca cultivation (as indicated in Table 3), particularly in steep terrains. Swift restorative measures are imperative to reinstate ecosystem services provided by grasslands. Despite the inclusion of high Andean natural pasture management for greenhouse gas reduction within the National Determined Contribution



(NDC), outlined by the Peruvian multisectoral working group (MINAM 2019), further measures are warranted to ensure the preservation of soil C stored within grasslands and unique “bofedales” ecosystems. Economic and social incentives for pastoralists must be implemented to guarantee the establishment of best management practices (rotational grazing, improved fallows with legumes, water harvesting, wetland, and grassland restoration) to avoid the expansion of the agricultural frontier. Special attention must be provided to “bofedales” which occupy around 0.8% of the Peru surface (~ 1.05 Mha) and are found predominantly in mine concessions (41% of total “bofedales” surface), keeping 21% of them under the custody of rural inhabitants (Fuentealba and Rios 2023). Despite that, an increase of + 2% year<sup>-1</sup> in areas of “bofedales” (by greater availability of water resources in dry seasons due to deglaciation) has been reported for the 1986–2005 period in the southern Andes (Pauca-Tanco and others 2020), in recent years there has been a reduction of areas of “bofedales.” Thus, some studies reported an area loss rate of – 3.8 to – 0.4% year<sup>-1</sup> during the 2005–2016 period (Machuca-Crespo 2018; Pauca-Tanco and others 2020; Pamo-Sedano and Oscco-Coa 2022). These ecosystems can be restored by establishing artificial “bofedales,” which can preserve the same ecosystem services as natural ones, as was remarked in recent studies (Monge-Salazar and others 2022).

The present study was conducted in the Junin National Reserve, which covers 5303.9 and 3608.8 ha of “bofedales” of the Junin and Pasco departments, respectively (Fuentealba and Rios 2023). Conservation areas can be crucial as a life lab to test and monitor restoration activities involving local communities, thus improving the geospatial modeling of SOC to build an interoperable public digital infrastructure that can serve as a monitoring-verification system for future compensation schemes for the benefit of indigenous pastoralists and rural inhabitants. Focusing on the ecologically significant and delineated regions of the Junin National Reserve and its buffer zone, our predictive mapping depicted distinct variations in SOC distribution. Specifically, within the reserve itself, approximately 32% of the C storing surfaces had SOC values over 9.6%, compared to only 8% within its buffer zone (Figure 5). While RSOC and  $\delta^{13}\text{C}_{\text{SOC}}$  are key variables that provide valuable information, the significant importance of pH—a site-specific sampled predictor—in their models limited our ability to produce accurate spatial distribution maps.

## CONCLUSION

Processes that drive SOC and fractions like RSOC and  $\delta^{13}\text{C}_{\text{SOC}}$  in high Andean rangeland systems have not been studied yet, challenging the choice of environmental predictors (LULC identification and classification, remote sensing products, climate and soil variables, among others) for their modeling. Under this context, ML algorithms capture nonlinear interaction and process complexity to model the studied soil C target variables with acceptable and consistent performance. “Bofedales” were the most important reservoirs in terms of the total and the refractory fraction of SOC compared to the other land uses. Its highest depletion of  $^{13}\delta\text{C}$  is a potential indicator of higher turnover rates, high plant productivity, and C persistence. Because “bofedales” are affected by strong perturbations (extraction of vegetation and part of the topsoil—“champeo,” overgrazing) in the study area, it is recommended to establish restoration activities to guarantee ecosystem services from those ecosystems. For example, the management of natural grasslands through cultivated pastures showed indicators of higher productivity (more depletion of  $\delta^{13}\text{C}$ ), remarking its potential for grassland restoration after crop encroachment (like maca crop) in this area.

Free, publicly available remote sensing data can be beneficial for SOC prediction. Vegetation indices close to the NIR band, such as NDMI and SER2, were good environmental predictors for the total soil C (SOC and  $\delta^{13}\text{C}_{\text{SOC}}$ ). However, to improve the prediction, vegetation and climatic indices must be complemented with data taken in situ, such as pH, and especially LULC, because it is the primary driver of SOC variation. Together, these variables can explain SOC dynamics, facilitating their prediction using ML algorithms. Considering the high reservoirs of C in the soils of highland Andean ecosystems, future SOC and fractions mapping will be essential for decision-makers and regional governments for compensation schemes in voluntary or regulated C markets. The SOC map elaborated in this study can be used for this aim, and some improvements can be achieved if more soil samplings are collected, especially in “bofedales,” improved and cultivated pastures, and fallows LULC.

## ACKNOWLEDGEMENTS

This research was carried out under the support of the “Innovate” Work Package of “Excellence in Agronomy” and the 3rd Work Package of “AgriLAC Resiliente” OneCGIAR initiatives. MC received

support from the USDA Foreign Agricultural Service, Borlaug Fellowship Program (BF-CR-16-009). The authors greatly thank producers of the study area in Junin, Peru, for allowing us to sample soils for their properties, Lake Junin National Reserve staff, and especially Alan Chamorro of ECOAN, for the help provided during fieldwork.

## OPEN ACCESS

This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## DATA AVAILABILITY

The data that support the findings of this study are openly available in Dataverse CGIAR repository at <https://doi.org/https://doi.org/10.21223/VVIFL9>.

## REFERENCES

- Adi SH, Grunwald S. 2020. Integrative environmental modeling of soil carbon fractions based on a new latent variable model approach. *Science of the Total Environment* 711:134566.
- Adler PB, Morales JM. 1999. Influence of environmental factors and sheep grazing on an Andean grassland. *Journal of Range Management* 52:471–481. <https://doi.org/10.2307/4003774>.
- Al-Abbas AH, Swain PH, Baumgardner MF. 1972. Relating organic matter and clay content to the multispectral radiance of soils. *Soil Science* 114(6):477–485.
- Alavi-Murillo G, Diels J, Gilles J, Willems P. 2022. Soil organic carbon in Andean high-mountain ecosystems: importance, challenges, and opportunities for carbon sequestration. *Regional Environmental Change* 22:128.
- Arce A, de Haan S, Juarez H, Dhar Burra D, Plasencia F, Ccanto R, Polreich S, Scurrah M. 2019. The Spatial-Temporal Dynamics of Potato Agrobiodiversity in the Highlands of Central Peru: A Case Study of Smallholder Management Across Farming Landscapes. *Land* 8:169. <https://doi.org/10.3390/land8110169>.
- Ayala Izurieta JE, Márquez CO, García VJ, Jara Santillán CA, Sisti JM, Pasqualotto N, Van Wittenberghe S, Delegido J. 2021. Multi-predictor mapping of soil organic carbon in the alpine tundra: a case study for the central Ecuadorian páramo. *Carbon Balance and Management* 16(1):1–19.
- Balabaeva K, Kovalchuk S. 2019. Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. *Procedia Computer Science* 156:87–96. <https://doi.org/10.1016/j.procs.2019.08.183>
- Bernoux M, Cerri CC, Neill C, de Moraes JF. 1998. The use of stable carbon isotopes for estimating soil organic matter turnover rates. *Geoderma* 82(1–3):43–58.
- Bishop JL, Lane MD, Dyar MD, Brown AJ. 2008. Reflectance and emission spectroscopy study of four groups of phyllosilicates: Smectites, kaolinite-serpentines, chlorites and micas. *Clay Minerals* 43:35–54.
- Caro C, Quinteros Z, Mendoza V. 2007. Identificación de indicadores de conservación para la Reserva Nacional de Junín. *Perú. Ecología Aplicada* 6(1–2):67–74.
- Caro C, Sánchez E, Quinteros Z, Castañeda L. 2014. Respuesta de los pastizales altoandinos a la perturbación generada por extracción mediante la actividad de “champeo” en los terrenos de la Comunidad Campesina Villa de Junín. *Perú. Ecología Aplicada* 13(2):85–95.
- Carré F, McBratney AB, Minasny B. 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma* 141(1–2):1–14.
- Catorci A, Cesaretti S, Velasquez JL, Malatesta L, Zeballos H. 2014. The interplay of land forms and disturbance intensity drive the floristic and functional changes in the dry Puna pastoral systems (southern Peruvian Andes). *Plant Biosystems* 148:547–557. <https://doi.org/10.1080/11263504.2014.900126>.
- Chatterjee S, Hartemink AE, Triantafyllis J, Desai AR, Soldat D, Zhu J, Townsend PA, Zhang Y, Huang J. 2021. Characterization of field-scale soil variation using a stepwise multi-sensor fusion approach and a cost-benefit analysis. *Catena* 201:105190.
- Chen S, Arrouays D, Mulder VL, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J, Meersmans J, Richer-de-Forges A, Walter C. 2022. Digital mapping of globalsoilmap soil properties at a broad scale: a review. *Geoderma* 409:115567.
- Cochi Machaca N, Condori B, Rojas Pardo A, Anthelme F, Meneses RI, Weeda CE, Perotto-Baldivieso HL. 2018. Effects of grazing pressure on plant species composition and water presence on bofedales in the Andes mountain range of Bolivia. *Mires Peat* 21:1–15.
- Cooper D, Wolf E, Colson C, Vering W, Granda A, Meyer M. 2010. Alpine Peatlands of the Andes, Cajamarca, Peru. *Arctic, Antarctic, and Alpine Research* 42:19–33.
- Corvalán C, Hales S, McMichael AJ. 2005. Ecosystems and human well-being: health synthesis, Millennium ecosystem assessment. *Millennium Ecosystem Assessment (Program)*, World Health Organization (Eds.). World Health Organization, Geneva, Switzerland.
- Dangles O, Carpio C, Barragan AR, Zeddard JL, Silvain JF. 2008. Temperature as a key driver of ecological sorting among invasive pest species in the tropical Andes. *Ecological Applications* 18:1795–1809.

- Davy MC, Koen TB. 2013. Variations in soil organic carbon for two soil types and six land uses in the Murray catchment, New South Wales, Australia. *Soil Research* 51(8):631–644.
- Diez A. 2020. Reforma agraria y procesos comunales: las comunidades de las SAIS Cahuide y Túpac Amaru en la sierra central del Perú. *Revista Del Instituto Riva-Agüero* 5:299–337. <https://doi.org/10.18800/revistaira.202002.010>.
- Dynarski KA, Bossio DA, Scow KM. 2020. Dynamic stability of soil carbon: reassessing the “permanence” of soil carbon sequestration. *Frontiers in Environmental Science* 8:514701.
- Ehleringer JR, Buchmann N, Flanagan LB. 2000. Carbon isotope ratios in belowground carbon cycle processes. *Ecological Applications* 10(2):412–422.
- Emadi M, Taghizadeh-Mehrjardi R, Cherati A, Danesh M, Movsavi A, Scholten T. 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing* 12(14):2234.
- Farley KA, Bremer LL, Harden CP, Hartsig J. 2013. Changes in carbon storage under alternative land uses in biodiverse Andean grasslands: implications for payment for ecosystem services. *Conservation Letters* 6:21–27.
- Farquhar GD, Ehleringer JR, Hubick KT. 1989. Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* 40:503–537.
- Fuentealba B, Ríos R. 2023. Memoria descriptiva inventario nacional de bofedales 2023. Instituto Nacional de Investigación en Glaciares y Ecosistemas de Montaña (INAIGEM). Huaraz, 205 p. <https://repositorio.inaigem.gob.pe/handle/16072021/466>
- Gamboa C, Bojacá CR, Schrevels E, Maertens M. 2020. Sustainability of smallholder quinoa production in the Peruvian Andes. *J. Clean. Prod.* 264:121657. <https://doi.org/10.1016/j.jclepro.2020.121657>.
- Gehl RJ, Rice CW. 2007. Emerging technologies for *in situ* measurement of soil carbon. *Climatic Change* 80(1–2):43–54.
- Gibbon A, Silman MR, Malhi Y, Fisher JB, Meir P, Zimmermann M, Dargie GC, Farfan WR, Garcia KC. 2010. Ecosystem carbon storage across the grassland-forest transition in the high Andes of Manu National Park, Peru. *Ecosystems* 13(7):1097–1111.
- Grunwald S. 2022. Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils. *Carbon Footprints* 1:5.
- Han R, Zhang Q, Xu Z. 2023. Responses of soil organic carbon cycle to land degradation by isotopically tracing in a typical karst area, southwest China. *PeerJ* 11:e15249. <https://doi.org/10.7717/peerj.15249>
- Hribljan JA, Suárez E, Heckman KA, Lilleskov EA, Chimner RA. 2016. Peatland carbon stocks and accumulation rates in the Ecuadorian páramo. *Wetlands Ecology and Management* 24:113–127.
- Jagadamma S, Lal R, Ussiri DA, Trumbore SE, Mestelan S. 2010. Evaluation of structural chemistry and isotopic signatures of refractory soil organic carbon fraction isolated by wet oxidation methods. *Biogeochemistry* 98:29–44.
- John K, Abraham Isong I, Michael Kebonye N, Okon Ayito E, Chapman Agyeman P, Marcus Afu S. 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* 9(12):487.
- Keskin H, Grunwald S, Harris WG. 2019. Digital mapping of soil carbon fractions with machine learning. *Geoderma* 339:40–58.
- Klink S, Keller AB, Wild AJ, Baumert VL, Gube M, Lehdorff E, Meyer N, Mueller CW, Phillips RP, Pausch J. 2022. Stable isotopes reveal that fungal residues contribute more to mineral-associated organic matter pools than plant residues. *Soil Biology and Biochemistry* 168:108634. <https://doi.org/10.1016/j.soilbio.2022.108634>
- Krull ES, Baldock JA, Skjemstad JO. 2003. Importance of mechanisms and processes of the stabilisation of soil organic matter for modelling carbon turnover. *Functional Plant Biology* 30(2):207–222.
- Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. 2019. Caret: classification and regression training. R Package Version 6:86.
- Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36:1–13.
- Lamichhane S, Kumar L, Wilson B. 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. *Geoderma* 352:395–413.
- Liu S, An N, Yang J, Dong S, Wang C, Yin Y. 2015a. Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China. *Catena* 133:137–144.
- Liu Y, Guo L, Jiang Q, Zhang H, Chen Y. 2015b. Comparing geospatial techniques to predict SOC stocks. *Soil and Tillage Research* 148:46–58.
- Liu D, Yu Z, Lin J. 2018. Application of combustion module coupled with cavity ring-down spectroscopy for simultaneous measurement of SOC and  $\delta^{13}\text{C}$ -SOC. *Journal of Spectroscopy* 2018:1–5.
- Liu X, Chen D, Yang T, Huang F, Fu S, Li L. 2020. Changes in soil labile and recalcitrant carbon pools after land-use change in a semi-arid agro-pastoral ecotone in Central Asia. *Ecological Indicators* 110:105925.
- Luo Z, Feng W, Luo Y, Baldock J, Wang E. 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. *Global Change Biology* 23(10):4430–4439.
- Machuca-Crespo DV. 2018. Efectos de la extracción de turba en un sistema socio-ecológico altoandino: bofedales de Carampoma-Lima (Bachelor’s thesis, Pontificia Universidad Católica del Perú).
- Maldonado F. 2014. An introduction to the bofedales of the Peruvian High Andes. *Mires and Peat* 15(5):1–13.
- Mantas V, Caro C. 2023. User-relevant land cover products for informed decision-making in the complex terrain of the Peruvian Andes. *Remote Sensing* 15(13):3303.
- McKay MD, Beckman RJ, Conover WJ. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2):239–245.
- MINAM. 2015. Mapa nacional de cobertura vegetal: memoria descriptiva. Ministerio del Ambiente. Dirección General de Evaluación, Valoración y Financiamiento del Patrimonio Natural. Lima-Perú.
- MINAM. 2019. Catálogo de Medidas de Mitigación. Ministerio del Ambiente. Dirección General de Cambio Climático y Desertificación. Lima-Perú.
- Mirchooli F, Kiani-Harchegani M, Darvishan AK, Falahatkar S, Sadeghi SH. 2020. Spatial distribution dependency of soil organic carbon content to important environmental variables. *Ecological Indicators* 116:106473.



- Monge-Salazar MJ, Tovar C, Cuadros-Adriazola J, Baiker JR, Montesinos-Tubée DB, Bonnesoeur V, Antiporta J, Román-Dañobeytia F, Fuentealba B, Ochoa-Tocachi BF, Buytaert W. 2022. Ecohydrology and ecosystem services of a natural and an artificial bofedal wetland in the central Andes. *Science of the Total Environment* 838:155968.
- More SJ, Ravi V, Raju S. 2022. Carbon isotope discrimination studies in plants for abiotic stress. In: Shanker C, Anand A, Maheswari M, Eds. *Shanker AK, . Climate Change and Crop Stress: Molecules to ecosystems*. Academic Press International Publishing. pp 493–537.
- Neina D. 2019. The role of soil pH in plant nutrition and soil remediation. *Applied and environmental soil science* 2019(1):5794869. <https://doi.org/10.1155/2019/5794869>
- Padarian J, Minasny B, McBratney AB. 2019. Using deep learning for digital soil mapping. *Soil* 5(1):79–89.
- Padbhushan R, Kumar U, Sharma S, Rana DS, Kumar R, Kohli A, Kumari P, Parmar B, Kaviraj M, Kumar Sinha A, Annapurna K, Gupta VV. 2022. Impact of land-use changes on soil properties and carbon pools in India: a meta-analysis. *Frontiers in Environmental Science* 9:722.
- Pamo-Sedano J, Oscco-Coa CE. 2022. Análisis espacio temporal del bofedal de la comunidad de Ancomarca (Tacna-Perú) durante el período 1990–2021, con técnicas de teledetección. *Revista Ciencias Biológicas y Ambientales* 1(1):43–53. <http://doi.org/10.33326/29585309.2022.1.1587>
- Parsaie F, Farrokhian Firouzi A, Mousavi SR, Rahmani A, Sedri MH, Homae M. 2021. Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. *Environmental Monitoring and Assessment* 193(4):1–15.
- Pauca-Tanco A, Ramos-Mamani C, Luque-Fernández CR, Talavera-Delgado C, Villasante-Benavides JF, Quispe-Turpo JP, Villegas-Paredes L. 2020. Análisis espacio temporal y climático del humedal altoandino de Chalhuanca (Perú) durante el periodo 1986–2016. *Revista de Teledetección* 55:105–118. <https://doi.org/10.4995/raet.2020.13325>
- Peng Y, Xiong X, Adhikari K, Knadel M, Grunwald S, Greve MH. 2015. Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PloS One* 10(11):e0142295.
- Polk MH, Young KR, Cano A, León B. 2019. Vegetation of andean wetlands (bofedales) in huascarán national park. *Peru. Mires and Peat* 24(01):1–26.
- Poveda G, Espinoza JC, Zuluaga MD, Solman SA, Garreaud R, Van Oevelen PJ. 2020. High impact weather events in the Andes. *Frontiers in Earth Science* 8:162.
- Quinn P, Beven K, Chevallier P, Planchon O. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes* 5(1):59–79.
- R Core Team. 2022. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Rolando JL, Turin C, Ramírez DA, Mares V, Moneris J, Quiroz R. 2017a. Key ecosystem services and ecological intensification of agriculture in the tropical high-Andean Puna as affected by land-use and climate changes. *Agriculture, Ecosystems & Environment* 236:221–233.
- Rolando JL, Dubeux JC, Perez W, Ramirez DA, Turin C, Ruiz-Moreno M, Comerford NB, Mares V, Garcia S, Quiroz R. 2017b. Soil organic carbon stocks and fractionation under different land uses in the Peruvian high-Andean Puna. *Geoderma* 307:65–72.
- Rolando JL, Dubeux JCB Jr, Ramirez DA, Ruiz-Moreno M, Turin C, Mares V, Sollenberger LE, Quiroz R. 2018. Land Use Effects on soil fertility and nutrient cycling in the Peruvian high-Andean Puna grasslands. *Soil Science Society of America Journal* 82:463–474.
- RSIS. 2021. Ramsar Sites Information Service: Reserva Nacional de Junín. <https://rsis.ramsar.org/es/ris/882> Accessed 20 Apr 2023.
- Safanelli JL, Chabrilat S, Ben-Dor E, Demattê JA. 2020. Multispectral models from bare soil composites for mapping topsoil properties over Europe. *Remote Sensing* 12(9):1369.
- Sainepo BM, Gachene CK, Karuma A. 2018. Assessment of soil organic carbon fractions and carbon management index under different land use types in Olesharo Catchment, Narok County. Kenya. *Carbon Balance and Management* 13(1):1–9.
- Salvador F, Moneris J, Rochefort L. 2014. Peatlands of the Peruvian Puna ecoregion: types, characteristics and disturbance. *Mires Peat* 15:1–17.
- Segnini A, Posadas A, Quiroz R, Milori D, Saab SC, Neto LM, Vaz CMP. 2010. Spectroscopic assessment of soil organic matter in wetlands from the high Andes. *Soil Science Society of America Journal* 74(6):2246–2253.
- Segnini A, de Souza AA, Novotny EH, Milori D, da Silva WTL, Bonagamba TJ, Posadas A, Quiroz R. 2013. Characterization of peatland soils from the high Andes through <sup>13</sup>C nuclear magnetic resonance spectroscopy. *Soil Science Society of America Journal* 77(2):673–679.
- SENAMHI: Servicio Nacional de Meteorología e Hidrología del Perú. 2022. Mapa Climático del Perú. <https://www.senamhi.gob.pe/?p=mapa-climatico-del-peru>. Last accessed 15/12/2022.
- Shafizadeh-Moghadam H, Minaei F, Talebi-khiyavi H, Xu T, Homae M. 2022. Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. *Catena* 212:106077.
- Signorel A, Aho K, Alfons A, Anderegg N, Aragon T, Arppe A, and others. 2022. DescTools: Tools for Descriptive Statistics. R package version 0.99.47.
- Singh M, Sarkar B, Sarkar S, Churchman J, Bolan N, Mandal S, Menon M, Purakayastha TJ, Beerling DJ. 2018. Stabilization of soil organic carbon as influenced by clay mineralogy. In: Sparks DL, editor. *Advances in agronomy*. Academic Press International Publishing. pp 33–84.
- Skarbø K, VanderMolen K. 2016. Maize migration: key crop expands to higher altitudes under climate change in the Andes. *Climate and Development* 8:245–255.
- Smith P. 2008. Soil organic carbon dynamics and land-use change. In: Braimoh AK, Vlek PLG, Eds. *Land use and soil resources*. Dordrecht International Publishing: Springer. pp 9–22.
- Smith CJ, Chalk PM. 2021. Carbon ( $\delta^{13}\text{C}$ ) dynamics in agroecosystems under traditional and minimum tillage systems: a review. *Soil Research* 59(7):661–672.
- Somarathna PDSN, Malone BP, Minasny B. 2016. Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. *Geoderma Regional* 7(1):38–48.
- Somarathna PDSN, Minasny B, Malone BP. 2017. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. *Soil Science Society of America Journal* 81(6):1413–1426.

- Sreenivas K, Dadhwal VK, Kumar S, Harsha GS, Mitran T, Sujatha G, Rama Suresh GJ, Fyze MA, Ravisankar T. 2016. Digital mapping of soil organic and inorganic carbon status in India. *Geoderma* 269:160–173.
- Stein M. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29(2):143–151.
- Strobl C, Boulesteix AL, Zeileis A, Hothorn T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 8:1–21. <https://doi.org/10.1186/1471-2105-8-25>
- Tovar C, Duivenvoorden JF, Sánchez-Vega I, Seijmonsbergen AC. 2012. Recent changes in patch characteristics and plant communities in the jalca grasslands of the Peruvian Andes. *Biotropica* 44:321–330. <https://doi.org/10.1111/j.1744-7429.2011.00820.x>.
- Tovar C, Seijmonsbergen AC, Duivenvoorden JF. 2013. Monitoring land use and land cover change in mountain regions: An example in the Jalca grasslands of the Peruvian Andes. *Landsc. Urban Plan.* 112:40–49. <https://doi.org/10.1016/j.landurbplan.2012.12.003>.
- Turin C, Carbajal M, Zorogastúa P, Chamorro A, Quiroz R. 2018. El boom de la maca: transformando paisajes y sociedades rurales de la zona central altoandina del Perú. 17 Seminario Permanente de Investigación Agraria (SEPIA). Cajamarca Perú. 29–31 Ago 2017.
- United States Geological Survey: USGS 02323500 SuwanneeRiver Near Wilcox, Fla. 2020. [https://waterdata.usgs.gov/usa/nwis/uv?site\\_no=02323500](https://waterdata.usgs.gov/usa/nwis/uv?site_no=02323500). Last accessed 24/07/2022.
- Veronesi F, Schillaci C. 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. *Ecological Indicators* 101:1032–1044.
- Wang S, Fan J, Song M, Yu G, Zhou L, Liu J, Zhong H, Gao L, Hu Z, Wu W, Song T. 2013. Patterns of SOC and soil <sup>13</sup>C and their relations to climatic factors and soil characteristics on the Qinghai-Tibetan Plateau. *Plant and Soil* 363:243–255.
- Wang G, Jia Y, Li W. 2015. Effects of environmental and biotic factors on carbon isotopic fractionation during decomposition of soil organic matter. *Scientific Reports* 5(1):11043.
- Wang Y, Qi Q, Bao Z, Wu L, Geng Q, Wang J. 2022. A novel sampling design considering the local heterogeneity of soil for farm field-level mapping with multiple soil properties. *Precision Agriculture* 24:1–22.
- Xia S, Song Z, Wang Y, Wang W, Fu X, Singh BP, Kuzyakov Y, Wang H. 2021. Soil organic matter turnover depending on land use change: Coupling C/N ratios, δ<sup>13</sup>C, and lignin biomarkers. *Land Degradation & Development* 32(4):1591–1605. <https://doi.org/10.1002/ldr.3720>
- Xiao Y, Xue J, Zhang X, Wang N, Hong Y, Jiang Y, Zhou Y, Teng H, Hu B, Lugato E, Richer-de-Forges A, Arrouays D, Shi Z, Chen S. 2022. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. *Geoderma* 428:116208.
- Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, Comerford NB. 2014. Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling & Software* 57:202–215.
- Yang JM, Yang JY, Liu S, Hoogenboom G. 2014. An evaluation of the statistical methods for testing the performance of crop models with observed data. *Agricultural Systems* 127:81–89.
- Yates LA, Aandahl Z, Richards SA, Brook BW. 2022. Cross validation for model selection: a review with examples from ecology. *Ecological Monographs* 93(1):e1557.
- Zemp M, Huss M, Thibert E, Eckert N, McNabb R, Huber J, Barandun M, Machguth H, Nussbaumer SU, Gärtner-Roer I, Thomson L, Paul F, Maussion F, Kutuzov S, Cogley JG. 2019. Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016. *Nature* 568:382–386.
- Zeraatpisheh M, Garosi Y, Owliaie HR, Ayoubi S, Taghizadeh-Mehrjardi R, Scholten T, Xu M. 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena* 208:105723.
- Zhu C, Wei Y, Zhu F, Lu W, Fang Z, Li Z, Pan J. 2022. Digital mapping of soil organic carbon based on machine learning and regression kriging. *Sensors* 22:8997.
- Zimmermann M, Meir P, Silman MR, Fedders A, Gibbon A, Malhi Y, Urrego DH, Bush MB, Feeley KJ, Garcia KC, Dargie GC, Farfan WR, Goetz BP, Johnson WT, Kline KM, Modi AT, Rurau NMQ, Staudt BT, Zamora F. 2010. No Differences in soil carbon stocks across the tree line in the Peruvian Andes. *Ecosystems* 13:62–74.