



Article

Multi-Horizon Predictive Soil Mapping of Historical Soil Properties Using Remote Sensing Imagery

Preston T. Sorenson ^{1,*} , Jeremy Kiss ¹ , Angela K. Bedard-Haughn ¹ and Steve Shirliff ²

¹ Department of Soil Science, College of Agriculture and Bioresources, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada

² Department of Plant Sciences, College of Agriculture and Bioresources, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada

* Correspondence: preston.sorenson@usask.ca

Abstract: There is increasing demand for more detailed soil maps to support fine-scale land use planning, soil carbon management, and precision agriculture in Saskatchewan. Predictive soil mapping that incorporates a combination of environmental covariates provides a cost-effective tool for generating finer resolution soil maps. This study focused on mapping soil properties for multiple soil horizons in Saskatchewan using historical legacy soil data in combination with remote sensing band indices, bare soil composite imagery, climate data, and terrain attributes. Mapped soil properties included soil organic carbon content (SOC), total nitrogen, cation exchange capacity (CEC), electrical conductivity (EC), inorganic carbon (IOC), sand and clay content, and total profile soil organic carbon stocks. For each of these soil properties, a recursive feature elimination was undertaken to reduce the number of features in the overall model. This process involved iteratively removing features such that random forest out-of-bag error was minimized. Final random forest models were built for each property and evaluated using an independent test dataset. Overall, predictive models were successful for SOC ($R^2 = 0.71$), total nitrogen ($R^2 = 0.65$), CEC ($R^2 = 0.46$), sand content ($R^2 = 0.44$) and clay content ($R^2 = 0.55$). The methods used in this study enable mapping of a greater geographic region of Saskatchewan compared to those previously established that relied solely on bare soil composite imagery.

Keywords: Saskatchewan; predictive soil mapping; bare soil composite imagery; multi-temporal remote sensing; random forest



Citation: Sorenson, P.T.; Kiss, J.; Bedard-Haughn, A.K.; Shirliff, S. Multi-Horizon Predictive Soil Mapping of Historical Soil Properties Using Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 5803. <https://doi.org/10.3390/rs14225803>

Academic Editors: Ruxandra Vintila and Frank Veroustrate

Received: 27 September 2022

Accepted: 15 November 2022

Published: 17 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Extensive soil surveying and mapping took place historically in the Province of Saskatchewan, however, most of these mapping efforts ended by 1998. These historical 1:100,000 soil survey maps are easily accessible in the Saskatchewan Soil Information System digital platform (SKSIS.ca) [1]. The soil survey maps were widely used for agricultural management and land use planning in Saskatchewan throughout the late 20th century. There is now increasing demand for more detailed soil maps to support fine-tuned land use planning, soil carbon management, and precision agriculture. An extensive literature exists regarding using predictive soil mapping (PSM) to generate finer scale soil data [2]. Particularly, PSM using machine-learning tools in combination with a suite of environmental covariates has been an increasing focus of research around the world [3].

One promising tool for generating environmental covariate data for PSM is the use of multi-temporal remote sensing data. Recent developments in cloud computing (i.e., Google Earth Engine [4]) and data processing have enabled collections of multi-temporal data to be accessed and processed more easily. Multi-temporal statistics of remote sensing band indices such as normalized difference vegetation index (NDVI) have improved PSM work in the southwestern United States for mapping soil texture and coarse fragments [5]. Research

focused on soil organic carbon mapping at fine spatial scales has found multispectral time series to be a useful environmental covariate [6]. Multi-temporal statistics of remote sensing band indices have also improved PSM efforts for mapping soil properties other than carbon, such as sand content and cation exchange capacity in Iran [7].

The processing of multi-temporal remote sensing data enables the generation of bare soil environmental covariate data, typically referred to as bare soil composite imagery (BSCI) [8]. Bare soil composite imagery is particularly necessary for PSM in the Canadian Prairies, and other regions with extensive conservation tillage, because when dead plant residue cover exceeds 20 percent of a remotely sensed image pixel, then direct estimation of soil properties becomes difficult [9,10]. Vegetation and dead plant residue-free pixels are now rare in Saskatchewan due to extensive conservation tillage practices [11]. The availability of BSCI imagery is also dependent on bare soils being present, which is not the case for permanent pasture or forested land. With the use of historical remote sensing datasets such as Landsat 5, it is possible to generate historical BSCI for Saskatchewan, although present day BSCI cannot be generated [12]. Bare soil composite imagery has been used for a variety of projects including work in Germany [13], Brazil [14], and globally [15] for predicting soil properties such as soil organic carbon. Bare soil composite imagery has already been used successfully to generate historical soil carbon and clay content maps for regions of Saskatchewan [12].

Predictive soil mapping studies tend to focus on mapping soil properties as a two-dimensional surface by providing a single prediction per x and y coordinate-defined pixel. However, soil is a three-dimensional body, with variation in the third dimension: depth. The variation in soil properties by depth is largely dependent on soil horizons. To better account for soil variation in all spatial directions, three-dimensional machine learning approaches have been developed [16,17]. Three-dimensional predictive soil mapping includes depth as a model covariate, thereby training the model to account for multiple soil depths and enabling prediction at multiple soil depths [17]. This approach has the advantage of maximizing the total training dataset along with accounting for variation in depth profiles [18]. This approach has been used to predictively map soil organic carbon in three dimensions across Australia [19] and Canada [20].

The overall objective of this research was to use a horizon based three-dimensional predictive soil mapping approach to map the entire agricultural portion of the province of Saskatchewan for a range of soil properties. Previous work was restricted to only those regions where bare soil was present at some time, which left areas unmapped [12]. This study aimed to apply a process that included BSCI that was capable of mapping regions where bare soils never occurred. Specifically, using historical multi-temporal remote sensing data along with BSCI, including a range of remote sensing features and indices and determine which were most important for predicting the various soil properties in Saskatchewan's agricultural regions. Overall, the eventual use of the soil data generated from this project will be to provide baseline historic carbon data for the province of Saskatchewan for change over time comparisons, as well as finer spatial resolution maps of more stable soil properties such as soil texture for applications such as agricultural land management.

2. Materials and Methods

2.1. Study Area

The study area for this project included the entire agricultural region of the Canadian Province of Saskatchewan. The soils of Saskatchewan are characterized by development in glacial deposits because of the Pleistocene glaciation. Saskatchewan has a continental climate, with agriculture a dominant land use in the southern portion of the province [21]. Major crops include wheat, barley, Canola, lentils, field peas, and flax. A significant portion of the southern portion of the province is also used for pasture and forages [21]. The dominant soil type in this region are Chernozemic soils, with significant proportions of Solonchic and Luvisolic soils (Table 1).

Table 1. Soil orders present in the legacy soil survey data [22]. Soil orders are presented according to the Canadian System of Soil Classification [23], the World Reference Base [24], and the United States Department of Agriculture Soil Classification System [25].

Canadian System of Soil Science Classification	World Reference Base	United States Department of Agriculture Soil Classification System	<i>n</i>
Brunisol	Cambisol	Inceptisol	3
Chernozem	Kastanozem, Chernozem, Greyzem, Phaeozem	Borolls	398
Gleysol	Gleysol	Aqu suborders	7
Luvisol	Luvisol	Boralfs, Udalfs	66
Regosol	Regosol	Entisol	25
Solonetz	Solonetz	Natric Great Groups	73
Vertisol	Vertisol	Haplocryerts	5

2.2. Soil Data

Historical soil survey data was the sole point data source used in this study (Table 1) due to the limited amount of publicly available soil point data in Saskatchewan. The data was collected as part of historical soil mapping and survey efforts conducted during the 1990s and earlier [22]. The samples were collected to represent modal soils from the dominant soil types in a region and were collected by horizon with analyses performed on the homogenized horizon samples. Sample locations were estimated relative to landmarks by the soil surveyors, and the location accuracy is estimated as approximately ± 300 m. The soil point data was retrieved from the National Pedon Database [22]. The sampling locations were in modal slope positions and typically represent the dominant soil conditions for a quarter section. As a result, they represent the average conditions associated typically with mid slope positions.

As the samples were analyzed by horizon, with multiple A, B, or C horizons sometimes present, soil property weighted averages were calculated for each master horizon. The weighted average A, B, and C horizon values (including all A, B, and C horizons for the profile) were calculated for soil organic carbon, total nitrogen, cation exchange capacity, electrical conductivity, inorganic carbon, sand, and clay content (Table S1). The weighted averages of each parameter were calculated by multiplying soil property values by the proportion they accounted for of the entire master horizon type (A, B, or C) and then summing the resulting values for the master horizon type.

The total profile soil organic carbon stocks were also determined for each soil point. As bulk density values were not available for each sample, a pedotransfer function was generated to predict bulk density values using the entire National Pedon Database, including samples from outside Saskatchewan. Using the ranger package in R [26], a random forest model was generated to predict bulk density as a function of soil organic carbon, sand, and clay contents. The pedotransfer function had an R^2 of 0.52 and a root mean square error of 0.2 g cm^{-3} (Table 2), based on an independent validation test set generated with a 75–25 train-test split (further explained in the model validation section). Based on a root mean square error of 0.2 g cm^{-3} , there is an estimated median error of 3 Mg ha^{-1} in the soil organic carbon stock results due to the pedotransfer function. These predicted bulk density values were used to calculate total profile soil organic carbon stocks from soil organic carbon and horizon thicknesses, on a kilogram per square meter basis.

Table 2. Predictive model and null model results for each soil property. The null model RMSE values reflect the performance of a model where every soil property prediction value is equal to the mean property value for the entire training data set. It represents a base performance to compare the predictive model performance against.

Soil Property	Horizon	Null Model Root Mean Square Error	Predictive Model R ²	Predictive Model Root Mean Square Error	Predictive Model Concordance Correlation Coefficient	S
Soil Organic Carbon (%)	Overall	1.14	0.71	0.61	0.83	−0.07
	A	1.39	0.49	0.85	0.64	−0.11
	B	0.77	0.21	0.40	0.39	−0.06
	C	1.09	0.11	0.31	0.25	−0.01
Total Nitrogen (%)	Overall	0.10	0.65	0.06	0.76	0.00
	A	0.12	0.48	0.07	0.60	0.01
	B	0.07	0.25	0.05	0.39	0.00
	C	0.09	0.36	0.04	0.44	0.00
Inorganic Carbon (%)	Overall	8.09	0.65	4.79	0.79	−0.25
	A	5.66	0.11	3.07	0.29	−0.75
	B	7.08	0.16	5.67	0.25	−0.10
	C	10.88	0.56	5.42	0.73	0.18
Electrical Conductivity (dS m ^{−1})	Overall	2.50	0.36	2.18	0.57	−0.61
	A	1.94	0.08	1.67	0.23	−0.28
	B	1.94	0.26	1.67	0.47	−0.50
	C	3.40	0.34	3.00	0.53	−1.09
Cation Exchange Capacity (meq 100 g ^{−1})	Overall	11.27	0.46	8.18	0.63	−0.42
	A	12.09	0.47	8.80	0.63	−0.61
	B	10.12	0.41	7.81	0.61	−0.22
	C	11.37	0.36	7.81	0.51	−0.40
Clay (%)	Overall	15.62	0.55	10.47	0.70	−0.47
	A	14.36	0.65	8.23	0.76	−1.05
	B	15.77	0.50	11.10	0.67	0.43
	C	16.81	0.49	12.00	0.66	−0.68
Sand (%)	Overall	22.55	0.44	16.99	0.64	−0.61
	A	21.46	0.52	14.89	0.70	0.57
	B	22.45	0.44	17.03	0.64	−1.84
	C	23.82	0.37	19.07	0.58	−0.74
Horizon Thickness (%)	Overall	0.55	0.76	0.27	0.86	0.00
	A	0.38	0.06	0.10	0.21	−0.03
	B	0.30	0.06	0.23	0.21	−0.01
	C	0.80	0.66	0.39	0.79	0.03
Bulk Density (g cm ^{−3})	Overall	0.3	0.52	0.20	0.66	<0.01
Soil Organic Carbon Stock (kg m ^{−2})	Overall	5.83	0.27	4.84	0.47	0.31

The soil orders (based on the Canadian System of Soil Classification [23]) of the Saskatchewan portion of the National Pedon Database consisted primarily of Chernozemic soils (398 out of 577 total profiles), with the next most common soil order being Solonchaks with 73 profiles (Table 1), then Luvisols with 66 profiles, followed by Regosols with 25 profiles. The remainder of the soils included seven Gleysols, five Vertisols, and three Brunisols. The equivalent soil orders in the World Reference Base [24] and the United States Department of Agriculture Soil Classification System [25] are provided in Table 1.

2.3. Remote Sensing Data

Historical multi-temporal remote sensing data was retrieved using Google Earth Engine [4] for the agricultural regions of Saskatchewan (Figure 1). Landsat 5 Tier 1 surface reflectance data was acquired from 1984 to 1995, with several band indices calculated (Table S2). Data from this time period was used to correspond to the general time period that the soil data was collected. While the training data was collected largely in the 1980s and earlier, remote sensing data up to 1995 was included to ensure enough bare soil pixels were available for more reliable averages for the BSCI, as discussed in the BSCI section. The same date range was used for the other remote sensing indices. The median anthocyanin reflectance index (ARI), which is associated with anthocyanin pigments [27], was calculated for data from the months of July and August. The median canopy response salinity index (CRSI) for July and August was also determined [28], along with the median NDVI and median soil adjusted vegetation index (SAVI) for July and August and separately for September and October. Data from July and August was selected as this corresponds to peak photosynthetic activity in this region. Data from September and October was retrieved as the majority of arable cropland would be harvested during this time period, helping to separate soil properties that would be associated with different land use types such as wetlands or pastureland. The standard deviation of NDVI was determined using data from May to October, to account for the entire growing season. Each of these indices were calculated using all available pixels for the given time period after filtering the image collections to only include pixels with vegetation coverage (i.e., not bare soil) based on the methodology described in [12]. Only those pixels with NDVI values greater than 0.3, normalized difference index 7 (NDI7; normalized differences of Landsat 5 near-infrared and shortwave infrared 2 bands) values greater than zero, and normalized burn ratio (NBR²; normalized differences of Landsat 5 near-infrared and shortwave infrared 2 bands) values greater than 0.1 were included. Clouds, shadows, and low-quality pixels were filtered using the Landsat quality assessment band (pixel_qa). Each band was spatially filtered using a circular 10 × 10 median focal filter. This focal filtering was done to account for the ±300 m location uncertainty associated with the soil point data. The final filtered images were then aggregated to a spatial resolution of 100 m. Mapping with 3 × 3 median focal filtering and 30 m pixels produced significantly worse results in previous work that incorporated the National Pedon Database data [12]. The Google Earth Engine scripts for calculating remote sensing indices have been made available on GitHub [29].

To account for larger-scale climate-related factors, temperature and precipitation data was retrieved from the Copernicus Climate Change Service [30]. Daily mean average air temperature data was obtained for the months of May to September from 1979 to 2020. The mean temperature was calculated for each pixel in the data's native resolution (pixel size = 27,830 m). Average annual precipitation was also calculated for 1979 to 2020. The precipitation data included both rain and snowfall data, as all 12 months were included in the calculation. The coarse spatial resolution of the climate data meant it accounted for larger-scale climatic gradients. The resulting temperature and precipitation rasters were resampled to 100 m to match the band index rasters using bicubic interpolation. Bicubic interpolation was used over simpler interpolation methods, such as nearest neighbor, because nearest neighbor interpolation caused straight-line artifacts along the original pixel boundaries in the resulting predictive soil maps.

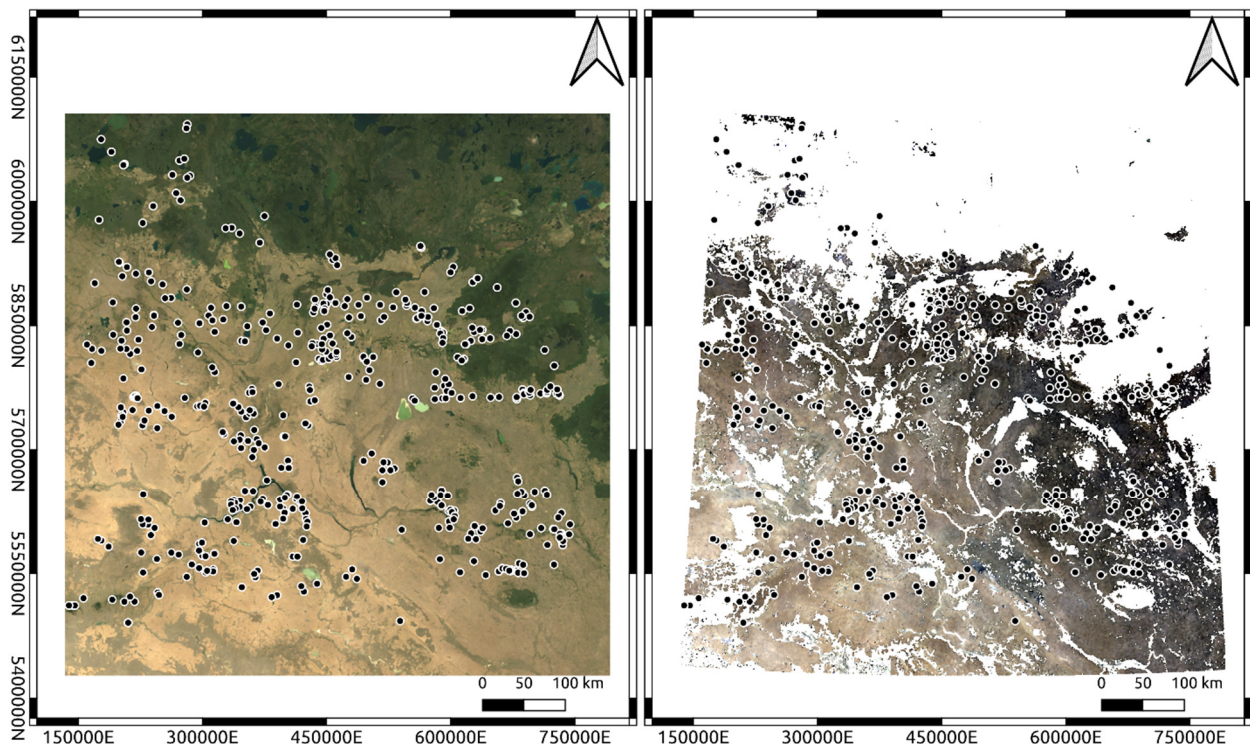


Figure 1. True colour red-green-blue median surface reflectance image (Landsat 5 red, green and blue bands) for Saskatchewan from May to October from 1984–1995 is on the left. True colour red-green-blue median reflectance bare soil pixel composite image (Landsat 5 red, green and blue bands) for Saskatchewan from 1984–1995 is on the right. The white areas correspond to areas either without bare soil pixels present or permanent pasture areas that have been masked. The black points indicate the location of the National Pedon Database sample locations used for model training and validation. Coordinates are in UTM Zone 13N NAD83.

2.4. Bare Soil Composite Imagery

Bare soil composite imagery was generated using the methodology described in [12] with Landsat 5 Tier 1 surface reflectance imagery for the months of July and August from 1984 to 1995. Data from this period was used to ensure that sufficient bare soil pixels with minimal crop residue were present in the imagery. The months of July and August were selected to increase the likelihood that crop residues from previous years had time to decompose. Clouds, shadows, and low-quality pixels were filtered using the Landsat quality assessment band (pixel_qa).

The image collection was filtered using NDVI, NBR_2 , NDI7, and the normalized difference water index (NDWI) calculations. Pixels were included as representing the bare soil surface if NDVI values were less than 0.3 [15], NBR_2 was less than 0.1 [15], NDWI was less than 0.5 [31], and NDI7 was less than 0 [32]. Areas that were mapped as pasture or grassland in Agriculture and Agri-Food Canada’s 2019 Annual Crop Inventory [21] were masked to further reduce the likelihood of non-bare soils being included in the final imagery. The bare soil imagery collections included data from much earlier than 2019; however, as both pasture and grasslands tend to be managed long-term, the use of this data for masking is considered appropriate. Once pixels in the collection had been filtered and masked, median reflectance values were calculated for each pixel. The resulting pixels were then spatially filtered using a circular 10×10 median focal filter. This focal filtering was done to account for ± 300 m location uncertainty associated with the soil point data. The final filtered composite image was then aggregated to a spatial resolution of 100 m. The BSCI is illustrated in Figure 1. Pixels that did not have bare soil present were assigned a null value of zero for model development. The Google Earth Engine used scripts for calculating BSCI have been made available on GitHub [33].

2.5. Terrain Attributes

Terrain attributes were derived from the 30 m digital surface model from the Advanced Land Observation Satellite (ALOS) [34]. Terrain attributes to account for local elevation variability and landscape-scale morphometric features were generated. Features related to elevation variability were the terrain ruggedness index (TRI) and the standard deviation of elevation. Multiple versions of these variables were generated by using a range of focal window sizes and elevation model median focal filtering sizes (Table S2). The input digital surface model was filtered using 3×3 , 5×5 , and 9×9 filter sizes. The TRI was calculated with window sizes of 10×10 and 20×20 , and standard deviation of elevation was calculated with 3×3 , 5×5 , 9×9 , 21×21 and 101×101 window sizes. These variables were selected as they represent coarser scale patterns in landscape variability and were hypothesized to be useful predictors given the relatively coarse scale of the elevation model relative to the short-range landscape variability of the prairies. Landscape-scale morphometric features, which included slope height, valley depth, mid slope position, normalized height, standardized height, and SAGA Wetness Index [35] were also included as they have been useful environmental covariates in other Saskatchewan PSM studies [36].

All TRI calculations were done using the system for automated geoscientific analyses (SAGA; SAGA Development Team 2011). The TRI values were calculated using either a window size of 10 or 20 pixels (Table S2) and were calculated for DEMs that had been filtered using median focal filtering with window sizes of either 3×3 , 5×5 , 9×9 , or not filtered. All standard deviation of elevation rasters were calculated in Google Earth Engine with window sizes of 3×3 , 5×5 , 9×9 , and 21×21 following 3×3 median focal filtering of the DEM. The standard deviation of elevation attribute was also calculated with window sizes of 21×21 and 101×101 following 9×9 median focal filtering of the DEM. The landscape-scale morphometric features were calculated using the Relative Heights and Slope Positions module in SAGA [35]. Module parameter values were left at default settings except for the t-value, which was set at 1000 to prioritize the importance of finer scale local depressions and peaks. All terrain attributes were aggregated to 100 m spatial resolution to match the resolution of the other covariates. It is important to note that for this region, slope positions often vary at finer scales than 30-m, and therefore so do important hydrological processes associated with relief that drive soil property variation. As a result, 30 m is often not fine enough resolution for landscape position scale mapping in this region.

2.6. Validation Data

To create an independent testing dataset for validation, the data was split into training and testing datasets using conditioned Latin Hypercube Sampling (cLHS) with the cLHS package in R [37]. A principal component analysis was performed on the entire list of environmental covariates (Table S2), and the first and second principal components were included as the inputs to the cLHS analysis. The cLHS was used to select 144 points (25 percent of the dataset) from the total dataset to be used as a testing dataset. The remaining 433 points (75 percent) were used as the training data for the study. The use of cLHS for creating the train-test split was to help ensure similar distributions of datasets across feature space [38]. The same training and testing datasets were used for each of the soil property models to remove differences in the training and testing data composition as a factor when comparing relative model performance. Profile observations with A-horizon total nitrogen values above 0.5 (95th percentile value was 0.43) were removed from the total nitrogen training and testing datasets. This was done to prevent rarely occurring outlier points from influencing the model, as they would unlikely be predicted reliably. Performance of the predictive model built for A-horizon total nitrogen without removing outliers had poor performance with an R^2 value of 0.09.

2.7. Model Development

The same predictive model development process was followed for each soil property, with recursive feature elimination as the first step. For each soil parameter, except for soil

organic carbon stocks, a three-dimensional predictive soil mapping approach was used [17]. In this study, values were predicted per soil horizon (A, B, and C) rather than for fixed depths. Soil horizon was included as a categorical covariate for the model to account for horizon variation in soil properties within one model. Profile soil organic carbon stocks were modelled without accounting for horizons. Predictive models were developed for the following soil properties:

- Soil organic carbon,
- Total nitrogen,
- Cation exchange capacity,
- Electrical conductivity,
- Inorganic carbon,
- Sand content,
- Clay content,
- Horizon thickness,
- Soil organic carbon stock

Using the ranger package in R [26], recursive feature elimination, using only the training data, was used to select the features to be included in the models for each soil property. The feature selection process was initially conducted separately on three groups of covariates: (1) BSCI bands, (2) band indices, and (3) terrain attributes to ensure that features from all three categories were included in the final models. First, a random forest model was built that included the BSCI visible light, near-infrared, and shortwave infrared bands, along with precipitation and temperature covariates to account for large-scale gradients. Feature importance was determined based on the variance of responses [26], and were then divided by the sum of the total feature importance values to make comparisons easier. The feature importance values indicate the relative effect a feature has on the overall model. Features were selected using a process where random forest models were iteratively built using sequentially less features based on their feature importance ranking. After each round, the feature importance values were determined and the least important feature was removed. The final BSCI features were selected based on where out-of-bag error was minimized across the iteratively built models. This process was repeated for the band indices and the terrain attribute groups of covariates.

Following feature selection per category of covariates, collinearly related features were managed: when two variables had a correlation above 0.9, only one was kept. The variable with the largest mean absolute correlation was then removed. Following recursive feature elimination for each covariate category, the variables selected for category were then subjected to a recursive feature elimination process was conducted that included all subsets of features from each category to select the features used in the final model. The list of features included in the final models for each soil property is listed in Table 3. Code for the feature selection process, model building, and mapping is available on GitHub [39]. Following feature selection, the ranger package in R [26] was used to build a final random forest model using the training dataset with 500 trees and the split rule as variance. A flow chart illustrating this process is provided in the supplemental material (Figure S1).

2.8. Model Validation

Model performance was validated using the independent test dataset which was withheld from the feature selection and model parameter optimization processes. The independent test dataset consisted of full profiles set aside from the training dataset so that predictions for individual horizons could not be informed by the adjacent horizons of the same profile. Values for the test dataset were predicted for each soil property and compared against the observed values. Performance was evaluated based on R^2 , root mean square error (RMSE), Lin's Concordance Correlation Coefficient (CCC), and Bias. Additionally, model performance was compared to a null model approach to evaluate model performance relative to an approach where no spatial pattern or environmental covariate relationships were modelled. This approach involved assigning the test dataset

soil property predictions values equal to the average soil property value of the entire training dataset (Table 2).

Table 3. Feature importance values for the final features included in each soil property model. The relative feature importance is the variance of responses for a given covariate by the sum of the all the variance of responses for all parameters in the model. Soil Adjusted Vegetation index is listed as SAVI, Canopy Response Salinity Index as CRSI, Normalized Difference Vegetation Index as NDVI, Anthocyanin Reflectance Index is ARI, and Topographic Roughness Index is TRI.

Soil Property	Features	Relative Feature Importance
Soil Organic Carbon (%)	Horizon	0.46
	ARI No Bare Soil Pixels	0.15
	Standard Deviation of NDVI	0.09
	September and October NDVI	0.08
	Precipitation	0.08
	Temperature	0.07
	Bare Soil Band 7	0.07
Bulk Density (g cm^{-3})	Soil Organic Carbon	0.26
	Sand Content	0.26
	Silt Content	0.25
	Clay Content	0.24
Profile Soil Organic Carbon Stocks (kg m^2)	Standard Deviation of NDVI	0.18
	Precipitation	0.14
	Temperature	0.14
	September and October NDVI	0.13
	CRSI No Bare Soil Pixels	0.12
	CRSI	0.12
	Bare Soil Band 2	0.11
	SAGA Wetness Index	0.05
Total Nitrogen (%)	Horizon	0.52
	Bare Soil Band 7	0.09
	September and October NDVI	0.08
	Standard Deviation of NDVI	0.08
	Temperature	0.08
	Precipitation	0.07
	CRSI No Bare Soil Pixels	0.07
Cation Exchange Capacity ($\text{meq } 100 \text{ g}^{-1}$)	Standard Deviation of NDVI	0.17
	Bare Soil Band 5	0.17
	Horizon	0.17
	July and August SAVI No Bare Soil Pixels	0.16
	Temperature	0.14
	Precipitation	0.13
	Standard Deviation of Elevation (101×101 focal window with 9×9 median focal filter of the input surface model)	0.06
Electrical Conductivity (dS m^{-1})	Horizon	0.18
	Temperature	0.16
	September and October NDVI	0.15
	CRSI	0.14
	Precipitation	0.12
	Bare Soil Band 5	0.12
	Standard Deviation of Elevation (21×21 focal window with 3×3 median focal filter of the input surface model)	0.07
	Standard Deviation of Elevation (3×3 focal window with 3×3 median focal filter of the input surface model)	0.06

Table 3. Cont.

Soil Property	Features	Relative Feature Importance
Inorganic Carbon (%)	Horizon	0.49
	Precipitation	0.21
	Temperature	0.11
	ARI No Bare Soil Pixels	0.10
	July and August NDVI	0.09
Clay (%)	Bare Soil Band 5	0.20
	Standard Deviation of NDVI	0.18
	September and October NDVI	0.17
	Temperature	0.13
	Precipitation	0.13
	CRSI No Bare Soil Pixels	0.13
Sand (%)	Horizon	0.05
	Standard Deviation of NDVI	0.20
	September and October of NDVI	0.17
	Bare Soil Band 7	0.15
	Temperature	0.14
	ARI No Bare Soil Pixels	0.13
	Precipitation	0.06
	Standardized Height	0.06
Horizon Thickness	Horizon	0.02
	Horizon	0.48
	Precipitation	0.14
	Temperature	0.13
	ARI	0.07
	Standard Deviation of NDVI	0.07
	Bare Soil Band 7	0.06
	Standard Deviation of Elevation (3 × 3 focal window with 3 × 3 median focal filter of the input surface model)	0.04

Performance of the models for each individual horizon was evaluated by separating the validation prediction results by horizon and calculating the model performance metrics separately, in addition to overall where values for all three horizons were included. This was done to determine how much of the overall model performance was driven by depth gradient trends present for some soil parameters. All predictions were determined using a single multi-horizon model for each soil parameter. Separate horizon specific models were not generated.

3. Results

The average historical A-horizon soil organic carbon contents ranged from 0 to 20.85% with median values of 2.30% (Table S1). Based on the independent validation data, the soil organic carbon model for all horizons had an R^2 value of 0.71 and an RMSE of 0.61% (Figure 2) compared to the null model RMSE value of 1.14% (Table 2). However, the model performance is variable across horizons, with A-horizon specific performance being lower at an R^2 of 0.49 (Table 2). The predictive model included seven features with horizon as the most important, followed by the ARI with bare soil pixels removed, standard deviation of NDVI and September and October median NDVI (Table 3). The performance for the profile soil organic carbon stock model was worse, with an R^2 of 0.27 and RMSE of 4.8 kg m⁻² (Figure 3) with the null model having an RMSE of 5.8 kg m⁻² (Table 2). The model included eight features in total, with standard deviation of NDVI, precipitation, temperature, and September and October NDVI as the most important features (Table 3).

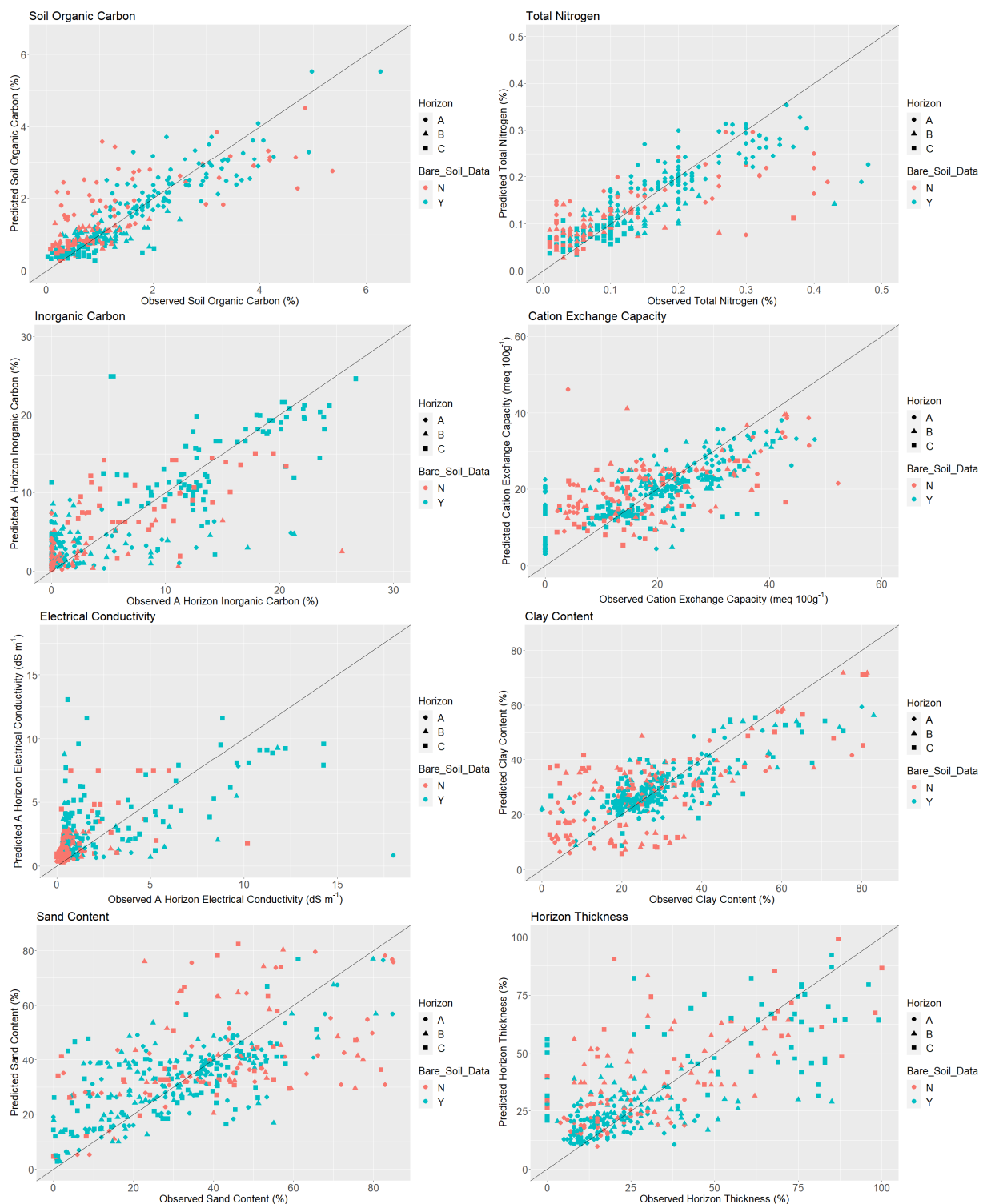


Figure 2. Average soil organic carbon (%), total nitrogen (%), inorganic carbon (%), cation exchange capacity (meq 100 g⁻¹), electrical conductivity (dS m⁻¹), clay content (%), sand content (%), and relative horizon thickness (%) predicted versus observed independent validation results. Model performance was assessed based on R^2 values, root mean square error (RMSE), Lin's Concordance Correlation Coefficient (pc) and bias. The solid black line indicates the 1:1 line to illustrate deviations between predicted and measured data. The Bare_Soil_Data legend indicates if bare soil composite imagery data was available for a point (Y) or not (N). The point shapes indicate which type of horizon the points are from.

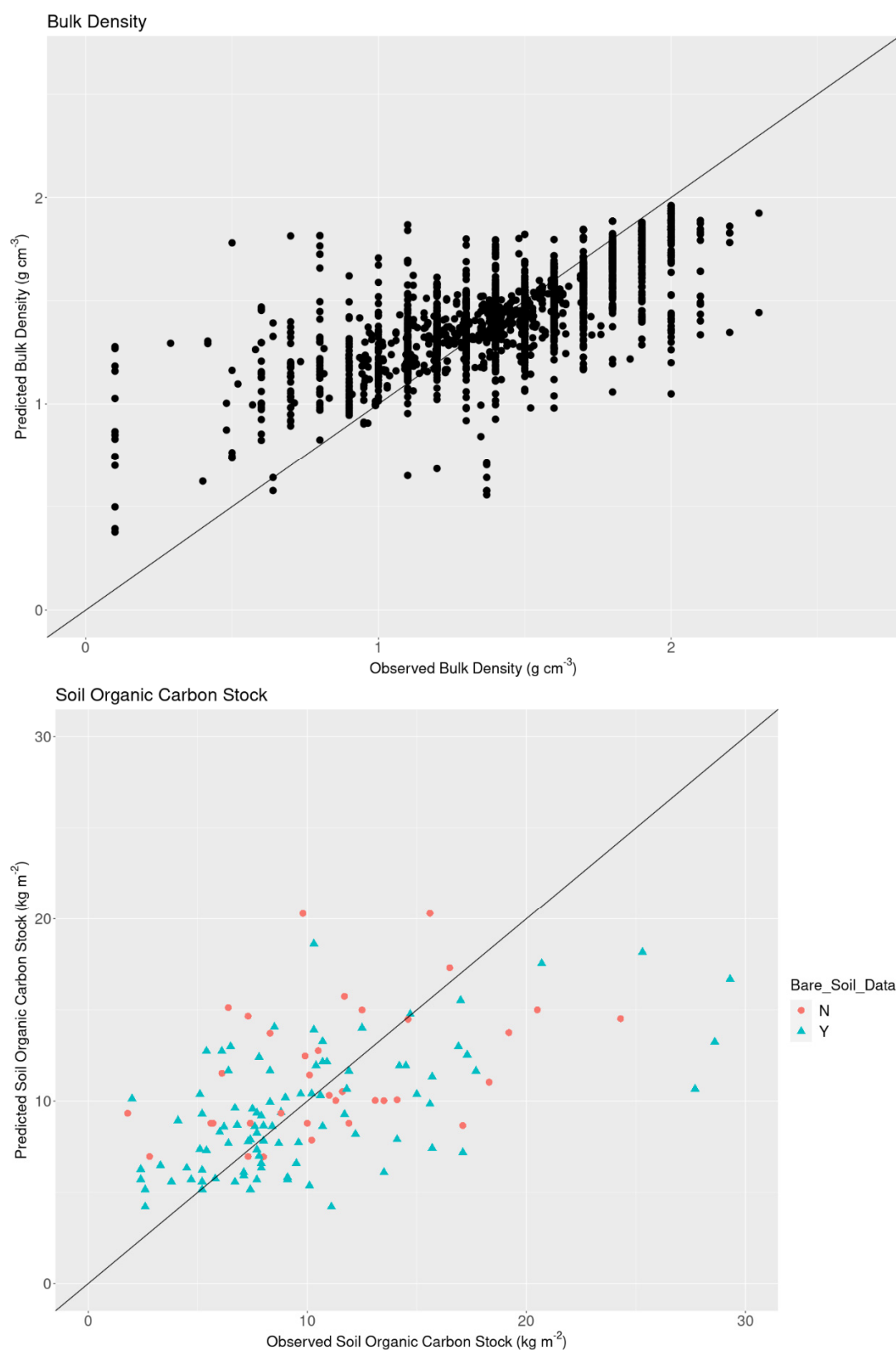


Figure 3. Bulk density (g cm^{-3}) and soil organic carbon stock (kg m^{-2}) predicted versus observed independent validation results. Model performance was assessed based on R^2 values, root mean square error (RMSE), Lin's Concordance Correlation Coefficient (ρ_c) and bias. The solid black line indicates the 1:1 line to illustrate deviations between predicted and measured data. The Bare_Soil_Data legend indicates if bare soil composite imagery data was available for a point (Y) or not (N).

Historical A-horizon total nitrogen values ranged from 0 to 2.72% with a median value of 0.20% (Table S1). The overall performance of the total nitrogen predictive model for all horizons based on the independent validation test was an R^2 of 0.65 and an RMSE of 0.06% (Figure 2). The null model had an RMSE of 0.10% (Table 2). As with the soil organic carbon model, part of the overall model performance was driven by average concentration differences amongst the soil horizons (Table 2). There were seven features included in the final A-horizon total nitrogen model: horizon, bare soil shortwave infrared two, September and October NDVI, standard deviation of NDVI, temperature, precipitation, and CRSI with bare soil pixels removed (Table 3). A-horizon cation exchange capacity ranged from 0 to 104.8 meq 100 g⁻¹ with a median value of 23.3 meq 100 g⁻¹ (Table S1). There were seven features in the final overall cation exchange capacity model: Standard Deviation of NDVI, bare soil shortwave infrared one, horizon, July and August SAVI with bare soil pixels removed, temperature, precipitation, and standard deviation nine by nine median focal filtered elevation with a 101 by 101 focal window (Table 3). The overall model performance based on independent validation was an R^2 of 0.46 and a RMSE of 8.18 meq 100 g⁻¹ (Figure 2). The cation exchange capacity null model had an RMSE of 11.27 meq 100 g⁻¹ (Table 2).

The A-horizon clay contents ranged from 0 to 80% with a median value of 23%, the A-horizon sand contents ranged from 0 to 95.8% with a median value of 35% (Table S1). The overall clay predictive model had an R^2 of 0.55 and an RMSE of 10.47% (Figure 2). The overall sand model had an R^2 of 0.44 and an RMSE of 16.99% (Figure 2). The null model RMSE values were 15.62% for clay and 22.55% for sand (Table 2). Model performance was better for the A-horizon compared to the overall model (Table 2). The clay model used seven features, including BSCI band 5, standard deviation of NDVI, September and October NDVI, precipitation, temperature, CRSI with bare soil pixels removed, and horizon (Table 3). The sand model differed with ARI compared to CRSI, BSCI band 7 instead of BSCI band 5, although the order of importance varied (Table 3).

Historical electrical conductivity values ranged from 0 to 18.0 dS m⁻¹ for the A-horizon, 0 to 24.9 dS m⁻¹ for the B-horizon, and 0 to 18.3 dS m⁻¹ for the C-horizon (Table S1). Overall, performance of the electrical conductivity predictive models was poor (Figure 2), with the model performing best for C horizons (Table 2). The electrical conductivity model had an R^2 value of 0.36 and an RMSE of 2.18 dS m⁻¹. For comparison the null model RMSE was 2.5 dS m⁻¹ (Table 2). The profile electrical conductivity model included eight features in the final model, with the most important features being horizon, temperature, September and October NDVI, and (Table 3).

Inorganic carbon concentrations in the A-horizon ranged from 0 to 29.5% with a median value of 0% (Table S1). For the B-horizon, the inorganic carbon values ranged from 0 to 53.20% with a median value of 0.7%, and 0 to 66.91% for the C-horizon (Table S1). Model performance for the model was an R^2 of 0.65 and RMSE of 4.79% (Figure 2). The null model RMSE was equal to the predictive model RMSE at 8.09% (Table 2). The majority of the performance was driven by the C-horizon values, with poor results for A and B horizons. This is likely because on average the A and B horizon inorganic carbon concentrations were low. The model only included five features, with horizon, precipitation, temperature, ARI with bare soil pixels removed, and July and August NDVI as the final selected features (Table 3).

4. Discussion

The results for historical soil organic carbon, clay content, and cation exchange capacity are similar to previously published results that used only BSCI as a predictor, and excluded areas without BSCI present [12]. The performance of the soil organic carbon and cation exchange capacity models were slightly poorer compared to those of the previous study (R^2 of 0.49 compared to 0.55 for A-horizon SOC and R^2 of 0.46 compared to 0.50 for A-horizon CEC). There was, however, an improvement in the clay model performance (R^2 of 0.65 for A-horizons compared to 0.44). The slight decrease in the soil organic carbon

and cation exchange capacity models R^2 values is likely a result of worse performance in locations where BSCI was not available (Figures 2 and 4). The clay content model may have performed better because coarser textured soil is often associated with non-arable cropland and remote sensing features that help characterize non-arable cropland were included in this study's models and were not included in those of the previous study (Figure 5). Clay content and sand content are typically inversely related in this region; soils with higher clay contents usually have lower sand contents and therefore, clay content can be related to the same remote sensing characteristics indicative of non-arable cropland.

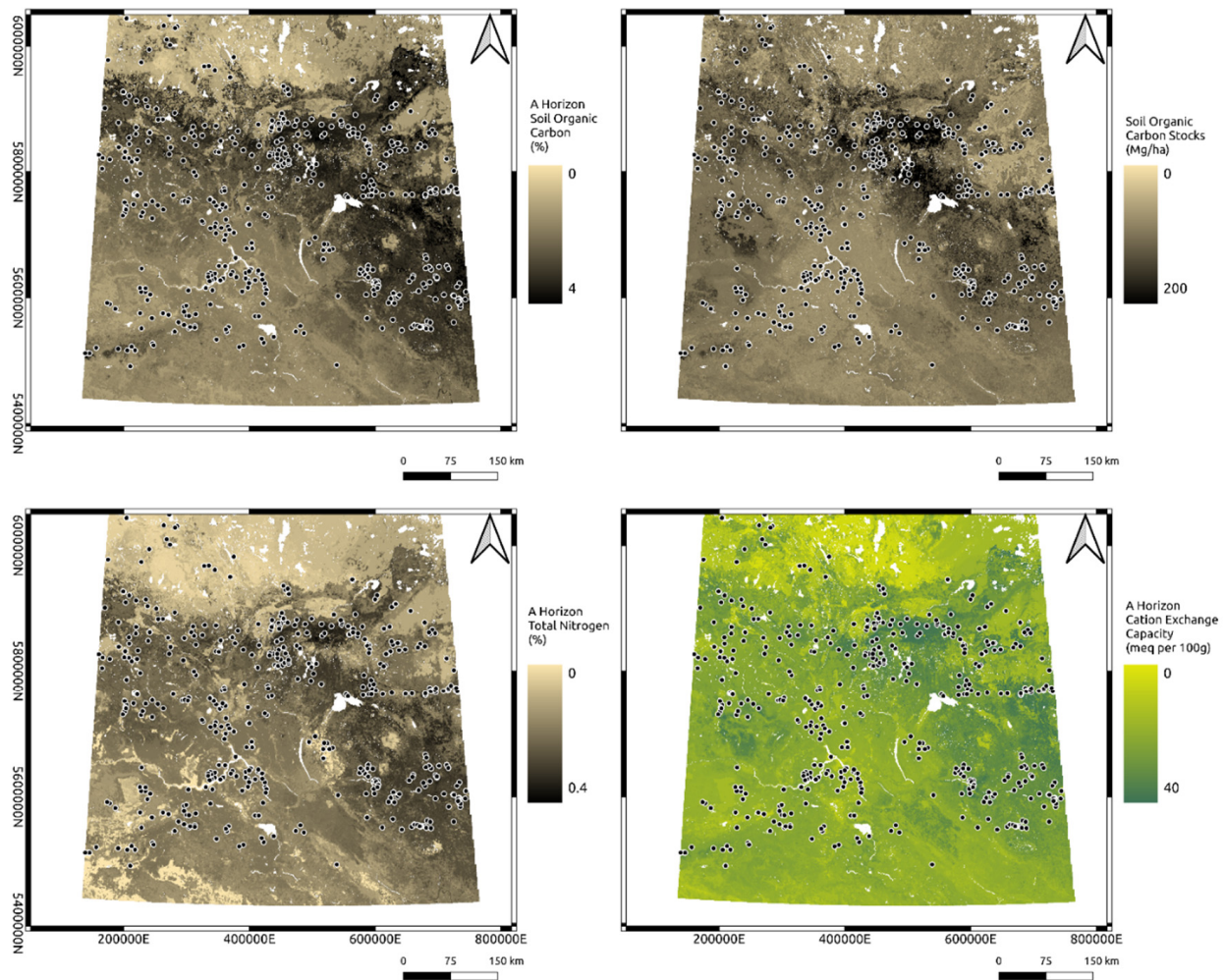


Figure 4. Maps for average A-horizon soil organic carbon (%), profile soil organic carbon stocks (Mg ha^{-1}), average A-horizon total nitrogen (%), and average A-horizon cation exchange capacity ($\text{meq } 100 \text{ g}^{-1}$) for Saskatchewan's agricultural region. White areas are masked water bodies. Coordinates are in UTM Zone 13N NAD83.

Overall, the results for soil organic carbon content models generated in this study compare reasonably well with other predictive mapping studies, particularly considering the $\pm 300 \text{ m}$ location uncertainty associated with the point data (Figure 4). Performance was similar to another Canadian study that mapped soil organic carbon concentrations near Ottawa, Ontario, which achieved an R^2 value of 0.46 and a CCC of 0.61 [40]. A previous study in France developed predictive models that achieved an R^2 of 0.59 for soil organic carbon content [41]. Whereas research focused on mapping soil organic matter on a smaller test area in Brazil had an R^2 of 0.38 [42], and ranged from 0.06 to 0.24 for a European-wide study [14].

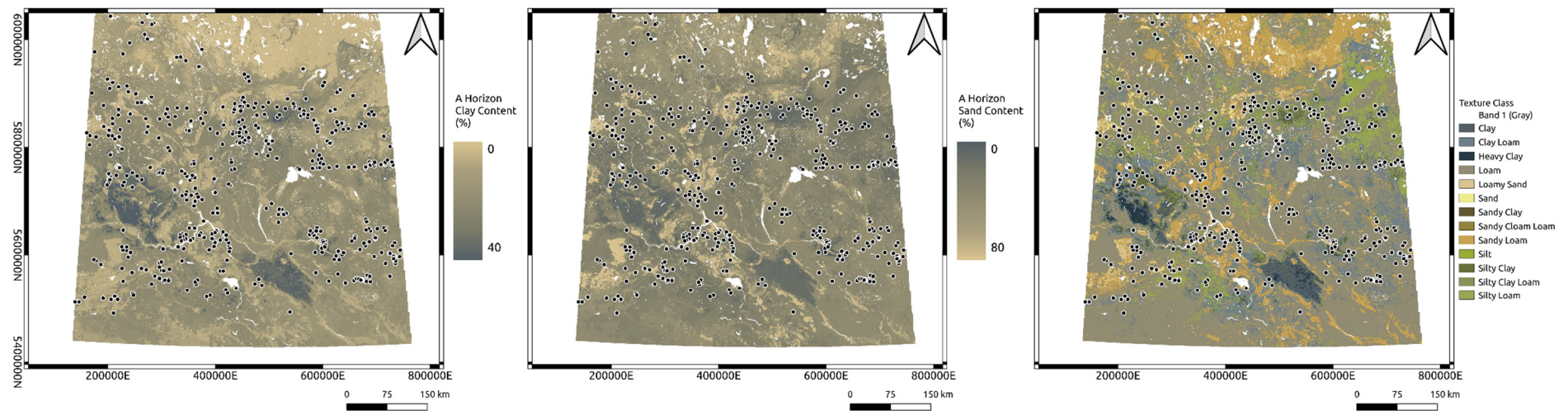


Figure 5. Maps for average A-horizon clay content (%), average A-horizon sand content (%) and A-horizon Canadian System of Soil Classification soil texture class for Saskatchewan's agricultural region. White areas are masked water bodies. Coordinates are in UTM Zone 13N NAD83.

The results of the historical profile soil organic carbon stock model were poorer than some other studies (R^2 of 0.27). Previous work predicting soil organic carbon stocks by using terrain, climate, and remote sensing variables and accurate spatial data in the Liaoning Province of China had R^2 values ranging from 0.6 to 0.7 [43]. A study in Australia on predicting soil organic carbon stocks in rangelands achieved R^2 values of 0.32 for 0–5 cm and 0.44 for 0–30 cm [44] and a study in South Australia’s agricultural zone achieved values between 0.40 and 0.45 [45]. The poorer results of the soil organic carbon stock model in this study could be due to the use of a pedotransfer function to estimate bulk density rather than direct measurements; that pedotransfer function achieved only an R^2 of 0.52 and a root mean square error of 0.2 (Figure 3). Additionally, accounting for the stocks to one meter may have also reduced accuracy due to the poorer performance of the soil organic carbon concentration model for B and C horizons (Table 2).

Models for other soil properties in this study achieved similar performance to those of other studies. In terms of clay content (R^2 values of 0.65 for A-horizon clay content in this study), other studies have achieved R^2 values of 0.34 [41], 0.23 to 0.37 [46], 0.62 [42], 0.44 [14] and 0.62 [47]. Cation exchange capacity model results reported in the literature have been quite variable with results ranging from being insufficient to report [41], to 0.24 [14], and 0.436 [46] compared to the R^2 value of 0.43 achieved in this study. Of note is that similar with this study, another study also had slightly lower results for predicting sand content compared to clay content [14]. This could be because the shortwave infrared BSCI directly responds to clay content, and not sand content, due to the location of clay shortwave infrared absorption features [48].

The results of the historical electrical conductivity predictive models in this study were poor (R^2 values ranging from 0.08 to 0.34 depending on horizon) and better results have been achieved in other studies. Research that has taken place in Iran achieved R^2 values of 0.57 for electrical conductivity and 0.54 for sodium adsorption ratio [49]. Additionally, other work in Iran had correlation coefficients ranging from 0.85 to 0.91 for predicting electrical conductivity [50]. The poor result for salinity in this study is not interpreted to be reflective of the overall potential for soil salinity mapping in the Canadian Prairies. The reason for the poor performance in this study can be attributed to the non-optimal training and validation dataset, specifically with regard to its limited number of profile observations with higher than average salinity values. The distribution of data is highly skewed towards low values, with a limited number of higher salinity data points (Figure 2). The sampling locations for the National Pedon Database were chosen to represent modal soil conditions, therefore there is underrepresentation of soils that typically occur as inclusions in larger soil polygons, such as saline soils in this region. The inclusion of synthetic aperture radar from Sentinel-1 may also help improve prediction results, as it has improved soil salinity mapping in other studies [51].

The historical profile electrical conductivity map appears to have significant areas where salinity values were overpredicted, which is not surprising given the large number of points in the salinity testing dataset that had low observed values but were predicted to have high values (Figure 6). While the model performance of the inorganic carbon model was adequate (Figure 2), the resulting map appears to have low utility (Figure 6). The map is largely a result of climatic gradients, with precipitation and temperature having a strong effect on the final model (Table 3). The utility of the map is low, especially as carbonates can vary at finer spatial scales in Saskatchewan than what will be captured in the map [52].

While the performance of the soil organic carbon and cation exchange capacity models was slightly worse than the BSCI only model [12], the approach in this study enables mapping of the entire agricultural area of Saskatchewan because it is not limited to only areas where BSCI is available. In total, 44 million hectares were mapped for Saskatchewan in this study. Mapping results for predictive models are presented in Figure 4 (A-horizon soil organic carbon, soil organic carbon stocks, A-horizon total nitrogen, and A-horizon cation exchange capacity), Figure 5 (A-horizon clay, A-horizon sand, and A-horizon texture class), and Figure 6 (C-horizon electrical conductivity and inorganic carbon). The scale

of the resulting mapping products needs to be interpreted with consideration of the focal filtering that was conducted prior to mapping. While variation at the 100 m pixel scale is evident in the maps, these trends have been smoothed by the 300×300 m filtering of the input data. This filtering was required given the uncertainty of the soil profile locations.

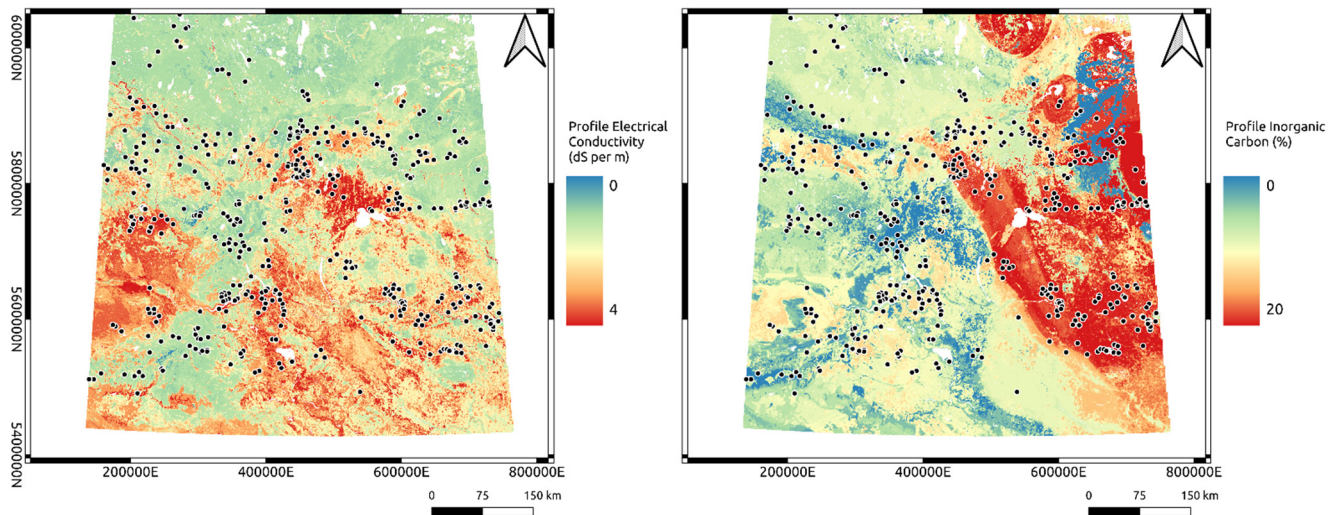


Figure 6. Maps for average C-horizon electrical conductivity (dS m^{-1}) and average C-horizon inorganic carbon content (%) for Saskatchewan's agricultural region. White areas are masked water bodies. Coordinates are in UTM Zone 13N NAD83.

Of note is that overall model performance for soil organic carbon and total nitrogen was influenced by the depth gradients present for some soil properties (Table 2). As there is a significantly higher concentration of soil organic carbon and total nitrogen in the A horizon as compared to the B and C horizon (Table S1), part of the overall model performance is driven by the models having different ranges of values to predict depending on soil horizon. For that reason, horizon by horizon comparisons are fairer for evaluating overall model utility for these parameters. On the contrary, the A horizon performance for the clay and sand model was higher (Table 2) than the overall model performance. This is likely because variations in B and C horizon soil textures was less closely correlated with environmental covariate data and no consistent depth trend was present, as observed with the soil organic carbon and total nitrogen data.

Several features were consistently important in the models for a range of soil properties. Of note is that models with poor performance and weaker relationships had more features included in the final model (Table 3), likely because there were no critical features that substantially improved performance during the recursive feature elimination. Overall, the climate variables of total precipitation and temperature were consistently important. Climate has long been understood to be an important soil forming factor [53]. These climatic variables influence the models for two reasons. The first reason is that climate affects plant growth and decomposition which in turn influences soil properties such as soil organic carbon directly. Indirectly, climate features help model performance by enabling the model to calibrate the influence of other covariates to climatic conditions. Of note is the low importance for many terrain attributes in the final models. This is likely a result of the available 30 m digital elevation model being too coarse to characterize landscape features in the study region. A higher resolution digital elevation model would likely improve model performance. Additionally, the 300 m location uncertainty associated with the training data is greater than the variation in slope positions, and as the training data typically was from modal slope positions data from upper and lower slope positions are underrepresented in the dataset.

In terms of remote sensing variables, September and October NDVI and Standard Deviation of NDVI were generally most important. The September and October NDVI vari-

able was important likely because arable cropland has very low NDVI by this time period, and so this feature helps to distinguish different longer term land uses, which is expected to be correlated with many soil properties. Additionally, while total photosynthetic activity is low in September and October in this region overall, the importance of the median September and October NDVI variable may be due to different NDVI responses from senescing vegetation in forests and grasslands. Standard deviation of NDVI has been shown to be a useful remote sensing variable for land use monitoring and land classification [54]; areas with higher NDVI standard deviations indicate more change in NDVI over the growing season. This is expected to be influenced by factors such as management practices, which in turn are influenced by soil and parent material type. Both standard deviation of NDVI and September and October NDVI variables are likely indirectly helping predict soil properties that are correlated with different land uses.

Bare soil composite imagery was a consistently important predictor as bare soil composite imagery bands were selected for the soil organic carbon, soil organic carbon stocks, total nitrogen, cation exchange capacity, electrical conductivity, clay content, sand content, and horizon thickness models. As the training data included points where bare soil imagery was not present, the absence of bare soil imagery itself is likely useful information as soil properties is tied to land use decisions, such as maintaining land as permanent pasture, in this region. The bare soil composite imagery near-infrared band was important for predicting soil organic carbon, likely because organic carbon has absorption features within the range of that band [48]. BSCI shortwave infrared one was important for soil organic carbon, total nitrogen, cation exchange capacity, and clay content. This is likely because soil organic carbon and clay have absorption features within the range of this band [48]. BSCI shortwave infrared two was important for the clay and sand models, as this band also includes clay absorption features [48]. However, given the bandwidths of the Landsat 5 bands are wide and include absorption features from a range of soil properties, it is difficult to conclusively decide if these are the reasons why the bands are important.

While less consistently important, July and August SAVI and the CRSI were also important for modelling cation exchange capacity (Table 3). This feature is highly correlated with leaf area index [55]. Cation exchange capacity is positively associated with crop yields [56] and the importance of this feature could be due to that relationship. The CRSI variable was important for a number of soil parameters. Canopy response salinity index has been documented to be related to salinity in other contexts [28]. It was important for the salinity models; however, it is difficult to draw conclusions about the feature importance in those models due to their overall poor performance. This feature is likely helping to predict clay content due to an overall correlation with plant activity, as clay content, at least in the Canadian Prairies, has been documented to be associated with increased yield [57].

Bare soil composite imagery has been shown to be a valuable dataset for mapping soil properties in contexts such as Europe [14] and in Brazil [42]. Recently it has also been used to produce maps of soil organic carbon, clay content, and cation exchange capacity in the Canadian Prairies [12]. However, one of the limitations of models that rely primarily on bare soil composite imagery is that they cannot be used to map soil properties in locations where bare soil was never exposed, such as the grasslands of the Canadian Prairies. Previous work has shown that inclusion of bare soil composite imagery with other datasets was useful for mapping soil properties in the middle east [47]. The results of this study indicate that the inclusion of bare soil imagery with other climate, remote sensing, and terrain attributes can be used to generate more extensive mapping in regions where bare soil pixels are not consistently present across the area of interest. Additionally, the bare soil composite imagery generation process [8] can be inverted to produce band ratio rasters that exclude bare soil pixels to help ensure the images were derived from pixels where vegetation is present, which also led to valuable predictors for historical soil property mapping in Saskatchewan.

The approach used in this study was successful for some, but not all, of the soil properties mapped in this study. The models for soil organic carbon, total nitrogen, cation

exchange capacity, clay content, and sand content were satisfactory. However, the electrical conductivity and inorganic carbon mapping was not successful in this study. Bare soil composite imagery was an important contributor to model success for each of the successful models, specifically the shortwave infrared bands. The bare soil composite imagery bands were not the most important bands for any of these variables. The overall model performance of the inorganic carbon model was satisfactory, however the artifacts in the final map made it unsuitable for use. While the bare soil composite imagery combined with other remote sensing variables did not lead to successful electrical conductivity and inorganic carbon maps, it is not fair to conclude these variables cannot be mapped. Potentially the poor model performance could be the result of training data values skewed towards low values. A training dataset with a higher number of samples with higher concentrations and a finer resolution digital elevation model could improve results.

5. Conclusions

Overall, historical soil maps were successfully generated for multiple soil properties including soil organic carbon, total nitrogen, cation exchange capacity, and sand and clay content. The resulting maps can be used to support change over time assessments of soil organic carbon concentrations in Saskatchewan. Additionally, the more detailed soil texture maps can be used to support applications such as hydrological modelling and agricultural land management. The mapping process used in this study will be applicable to the rest of the Canadian Prairies and the Northern Great Plains more broadly. The general process can be expected to be useable in other regions and climates, however additional or alternative remote sensing variables may need to be considered. While the predictive model performance was slightly lower for soil organic carbon and cation exchange capacity compared to previous work that used only BSCI data, the approach used in this study had the benefit of being capable of mapping Saskatchewan's entire agricultural region. Overall, the approach used in this study can be applied to areas where bare soils do not occur. However, the tradeoff is slightly lower accuracy and a less explainable model compared to using BSCI only. Generally, terrain attributes were not important features for the more successful models, which is likely due to the relatively coarse scale DEM available for Saskatchewan. Topographic variation occurs over fine spatial scales with low relief in the Saskatchewan prairies, which cannot be captured in the available coarse-scale DEMs. Terrain attributes derived from a finer resolution DEM would be more likely to improve predictive mapping efforts. The poor performance of the electrical conductivity and inorganic carbon models was likely due both to the coarse DEM used to generate terrain derivatives and because the range of values for these properties present within the training dataset did not encompass the range of values found on the landscape. Further work focused on modelling these soil properties using higher resolution DEMs is needed to identify successful approaches for predictively mapping them in Saskatchewan.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs14225803/s1>, Figure S1: Predictive soil mapping process diagram illustrating feature selection and model building process; Table S1: Soil property characteristics for the legacy soil survey data. The A-horizon values are the weighted average for all A-horizon values in a profile. Solum values are the weighted average for the A- and B-horizons, and the profile values are the weighted average for all horizons.; Table S2: Complete list of features included in the analysis prior to feature selection. Band ratio equations for each band ratio used as an environmental covariate in the analysis correspond to Landsat 5. Environmental covariate rasters are available on Zenodo at DOI: 10.5281/zenodo.7311804.

Author Contributions: Conceptualization, P.T.S. and A.K.B.-H.; methodology, P.T.S. and S.S.; software, P.T.S.; validation, P.T.S. and J.K.; formal analysis, P.T.S.; investigation, P.T.S.; resources, P.T.S.; data curation, P.T.S.; writing—original draft preparation, P.T.S.; writing—review and editing, J.K., A.K.B.-H. and S.S.; visualization, P.T.S.; supervision, A.K.B.-H.; project administration, A.K.B.-H.; funding acquisition, P.T.S. and A.K.B.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada, grant number PDF-557515-2021.

Data Availability Statement: National Pedon Database Data can be accessed here: <https://open.canada.ca/data/en/dataset/6457fad6-b6f5-47a3-9bd1-ad14aea4b9e0> (accessed on 14 November 2022). Google Earth Engine scripts for generating covariate data and R scripts for the predictive modelling are available at: <https://github.com/prestonsorenson> (accessed on 14 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bedard-Haughn, A.; Van Rees, K.; Bentham, M.; Krug, P.; Kiss, J.; Walters, K.; Heung, B.; Jamsrandorj, T.; Deters, R.; Cerkowniak, D.; et al. Saskatchewan Soil Information System (SKSIS) The Launch! Available online: <https://harvest.usask.ca/bitstream/handle/10388/8664/A.Bedard-Haughn%20et%20al.2018.pdf?sequence=1> (accessed on 3 February 2021).
2. Carré, F.; McBratney, A.B.; Mayr, T.; Montanarella, L. Digital Soil Assessments: Beyond DSM. *Geoderma* **2007**, *142*, 69–79. [CrossRef]
3. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
4. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
5. Maynard, J.J.; Levi, M.R. Hyper-Temporal Remote Sensing for Digital Soil Mapping: Characterizing Soil-Vegetation Response to Climatic Variability. *Geoderma* **2017**, *285*, 94–109. [CrossRef]
6. Guo, L.; Fu, P.; Shi, T.; Chen, Y.; Zhang, H.; Meng, R.; Wang, S. Mapping Field-Scale Soil Organic Carbon with Unmanned Aircraft System-Acquired Time Series Multispectral Images. *Soil Tillage Res.* **2020**, *196*, 104477. [CrossRef]
7. Fatholouloumi, S.; Vaezi, A.R.; Alavipanah, S.K.; Ghorbani, A.; Saurette, D.; Biswas, A. Improved Digital Soil Mapping with Multitemporal Remotely Sensed Satellite Data Fusion: A Case Study in Iran. *Sci. Total Environ.* **2020**, *721*, 137703. [CrossRef]
8. Demattê, J.A.M.; Fongaro, C.T.; Rizzo, R.; Safanelli, J.L. Geospatial Soil Sensing System (GEOS3): A Powerful Data Mining Procedure to Retrieve Soil Spectral Reflectance from Satellite Images. *Remote Sens. Environ.* **2018**, *212*, 161–175. [CrossRef]
9. Bartholomeus, H.; Epema, G.; Schaepman, M. Determining Iron Content in Mediterranean Soils in Partly Vegetated Areas, Using Spectral Reflectance and Imaging Spectroscopy. *Int. J. Appl. Earth Obs. Geoinf.* **2007**, *9*, 194–203. [CrossRef]
10. Bartholomeus, H.M.M.; Schaepman, M.E.E.; Kooistra, L.; Stevens, A.; Hoogmoed, W.B.B.; Spaargaren, O.S.P. Spectral Reflectance Based Indices for Soil Organic Carbon Quantification. *Geoderma* **2008**, *145*, 28–36. [CrossRef]
11. Statistics Canada. 2016 Census of Agriculture. Available online: <https://www.statcan.gc.ca/eng/ca2016> (accessed on 27 August 2020).
12. Sorenson, P.T.; Shirtliffe, S.J.; Bedard-Haughn, A.K. Predictive Soil Mapping Using Historic Bare Soil Composite Imagery and Legacy Soil Survey Data. *Geoderma* **2021**, *401*, 115316. [CrossRef]
13. Rogge, D.; Bauer, A.; Zeidler, J.; Mueller, A.; Esch, T.; Heiden, U. Building an Exposed Soil Composite Processor (SCMaP) for Mapping Spatial and Temporal Characteristics of Soils with Landsat Imagery (1984–2014). *Remote Sens. Environ.* **2018**, *205*, 1–17. [CrossRef]
14. Safanelli, J.L.; Chabrilat, S.; Ben-Dor, E.; Demattê, J.A.M. Multispectral Models from Bare Soil Composites for Mapping Topsoil Properties over Europe. *Remote Sens.* **2020**, *12*, 1369. [CrossRef]
15. Demattê, J.A.M.; Safanelli, J.L.; Poppiel, R.R.; Rizzo, R.; Silvero, N.E.Q.; de Sousa Mendes, W.; Bonfatti, B.R.; Dotto, A.C.; Salazar, D.F.U.; Alcântara de Oliveira Mello, F.; et al. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. *Sci. Rep.* **2020**, *10*, 4461. [CrossRef]
16. Hengl, T.; Macmillan, R.A. *Predictive Soil Mapping with R*; Open Geo Hub Foundation: Wageningen, The Netherlands, 2019; ISBN 978-0-359-30635-0.
17. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.M.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
18. Sanderman, J.; Hengl, T.; Fiske, G.; Solvik, K.; Adame, M.F.; Benson, L.; Bukoski, J.J.; Carnell, P.; Cifuentes-Jara, M.; Donato, D.; et al. A Global Map of Mangrove Forest Soil Carbon at 30 m Spatial Resolution. *Environ. Res. Lett.* **2018**, *13*, 055022. [CrossRef]
19. Viscarra Rossel, R.A.; Chen, C.; Grundy, M.J.; Searle, R.; Clifford, D.; Campbell, P.H. The Australian Three-Dimensional Soil Grid: Australia's Contribution to the GlobalSoilMap Project. *Soil Res.* **2015**, *53*, 845. [CrossRef]
20. Sothe, C.; Gonsamo, A.; Arabian, J.; Snider, J. Large Scale Mapping of Soil Organic Carbon Concentration with 3D Machine Learning and Satellite Observations. *Geoderma* **2022**, *405*, 115402. [CrossRef]
21. Agriculture and Agri-Food Canada. Annual Crop Inventory. Available online: <https://open.canada.ca/data/en/dataset/ba2645d5-4458-414d-b196-6303ac06c1c9> (accessed on 31 January 2021).
22. Agriculture and Agri-Food Canada. National Pedon Database. Available online: <https://open.canada.ca/data/en/dataset/6457fad6-b6f5-47a3-9bd1-ad14aea4b9e0> (accessed on 16 February 2021).
23. Soil Classification Working Group the Canadian System of Soil Classification. *The Canadian System of Soil Classification*, 3rd ed.; Agriculture and Agri-Food Canada: Ottawa, ON, Canada, 1998; Volume 187.

24. IUSS Working Group WRB. *World Reference Base for Soil Resources 2014: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*; Food and Agriculture Organization of the United Nations (FAO): Rome, Italy, 2014; ISBN 9789251083697.
25. Baillie, I.C. *Soil Survey Staff 1999, Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*; Agricultural Handbook 436; Natural Resources Conservation Service (USDA): Washington, DC, USA, 1999.
26. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [[CrossRef](#)]
27. Gitelson, A.A.; Merzlyak, M.N.; Chivkunova, O.B. Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves. *Photochem. Photobiol.* **2001**, *74*, 38. [[CrossRef](#)]
28. Scudiero, E.; Skaggs, T.H.; Corwin, D.L. Regional-Scale Soil Salinity Assessment Using Landsat ETM + Canopy Reflectance. *Remote Sens. Environ.* **2015**, *169*, 335–343. [[CrossRef](#)]
29. Sorenson, P. Google Earth Engine Scripts for Generating Predictive Soil Mapping Environmental Covariates. Available online: https://github.com/prestonsorenson/Google_Earth_Engine_PSM/tree/main (accessed on 16 August 2021).
30. Copernicus Climate Change Service (C3S). ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate. Available online: <https://cds.climate.copernicus.eu/cdsapp#!/home> (accessed on 14 October 2021).
31. Du, Y.; Zhang, Y.; Ling, F.; Wang, Q.; Li, W.; Li, X. Water Bodies' Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the SWIR Band. *Remote Sens.* **2016**, *8*, 354. [[CrossRef](#)]
32. Kokaly, R.F.; Clark, R.N.; Swayze, G.A.; Livo, K.E.; Hoefen, T.M.; Pearson, N.C.; Wise, R.A.; Benz, W.M.; Lowers, H.A.; Driscoll, R.L.; et al. *USGS Spectral Library Version 7*; United States Geological Survey (USGS): Reston, VA, USA, 2017; Volume 7.
33. Sorenson, P. Landsat 5 Bare Soil Composite Script. Available online: https://github.com/prestonsorenson/GEE_Bare_Soil_Composite/blob/main/Bare_Soil_Composite_Landsat_5 (accessed on 30 April 2021).
34. Japan Aerospace Exploration Agency. ALOS Global Digital Surface Model. Available online: <https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm> (accessed on 24 September 2021).
35. Boehner, J.; Selige, T. Spatial Prediction of Soil Attributes Using Terrain Analysis and Climate Regionalisation. In *SAGA—Analyses and Modelling Applications*; Boehner, J., McCloy, K.R., Strobl, J., Eds.; Göttinger Geographische Abhandlungen; Geographical Institute of the University of Göttingen: Göttingen, Germany, 2006; pp. 13–27.
36. Kiss, J. Predictive Mapping of Wetland Types and Associated Soils through Digital Elevation Model Analyses in the Canadian Prairie Pothole Region. M.Sc. Thesis, University of Saskatchewan, Saskatoon, SK, Canada, 2018.
37. Roudier, P. Clhs: A R Package for Conditioned Latin Hypercube Sampling 2011. Available online: <https://github.com/pierrerroudier/clhs/> (accessed on 14 November 2021).
38. Biswas, A.; Zhang, Y. Sampling Designs for Validating Digital Soil Maps: A Review. *Pedosphere* **2018**, *28*, 1–15. [[CrossRef](#)]
39. Sorenson, P. Predictive Soil Mapping Scripts. Available online: https://github.com/prestonsorenson/Predictive_Soil_Mapping (accessed on 28 October 2021).
40. Kasraei, B.; Heung, B.; Saurette, D.D.; Schmidt, M.G.; Bulmer, C.E.; Bethel, W. Quantile Regression as a Generic Approach for Estimating Uncertainty of Digital Soil Maps Produced from Machine-Learning. *Environ. Model. Softw.* **2021**, *144*, 105139. [[CrossRef](#)]
41. Vaysse, K.; Lagacherie, P. Evaluating Digital Soil Mapping Approaches for Mapping Global Soil Map Soil Properties from Legacy Data in Languedoc-Roussillon (France). *Geoderma Reg.* **2015**, *4*, 20–30. [[CrossRef](#)]
42. Rizzo, R.; Medeiros, L.G.; de Mello, D.C.; Marques, K.P.P.; de Mendes, W.S.; Quiñonez Silvero, N.E.; Dotto, A.C.; Bonfatti, B.R.; Demattê, J.A.M. Multi-Temporal Bare Surface Image Associated with Transfer Functions to Support Soil Classification and Mapping in Southeastern Brazil. *Geoderma* **2020**, *361*, 114018. [[CrossRef](#)]
43. Wang, S.; Zhuang, Q.; Wang, Q.; Jin, X.; Han, C. Mapping Stocks of Soil Organic Carbon and Soil Total Nitrogen in Liaoning Province of China. *Geoderma* **2017**, *305*, 250–263. [[CrossRef](#)]
44. Wang, B.; Waters, C.; Orgill, S.; Gray, J.; Cowie, A.; Clark, A.; Liu, D.L. High Resolution Mapping of Soil Organic Carbon Stocks Using Remote Sensing Variables in the Semi-Arid Rangelands of Eastern Australia. *Sci. Total Environ.* **2018**, *630*, 367–378. [[CrossRef](#)]
45. Liddicoat, C.; Maschmedt, D.; Clifford, D.; Searle, R.; Herrmann, T.; MacDonald, L.M.; Baldock, J. Predictive Mapping of Soil Organic Carbon Stocks in South Australia's Agricultural Zone. *Soil Res.* **2015**, *53*, 956–973. [[CrossRef](#)]
46. Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L.; Schaepman, M.E.; Papritz, A. Evaluation of Digital Soil Mapping Approaches with Large Sets of Environmental Covariates. *Soil* **2018**, *4*, 1–22. [[CrossRef](#)]
47. Poppiet, R.R.; Demattê, J.A.M.; Rosin, N.A.; Campos, L.R.; Tayebi, M.; Bonfatti, B.R.; Ayoubi, S.; Tajik, S.; Afshar, F.A.; Jafari, A.; et al. High Resolution Middle Eastern Soil Attributes Mapping via Open Data and Cloud Computing. *Geoderma* **2021**, *385*, 114890. [[CrossRef](#)]
48. Rossel, R.A.V.; Behrens, T.; Viscarra Rossel, R.A.; Behrens, T. Using Data Mining to Model and Interpret Soil Diffuse Reflectance Spectra. *Geoderma* **2010**, *158*, 46–54. [[CrossRef](#)]
49. Nabiollahi, K.; Taghizadeh-Mehrjardi, R.; Shahabi, A.; Heung, B.; Amirian-Chakan, A.; Davari, M.; Scholten, T. Assessing Agricultural Salt-Affected Land Using Digital Soil Mapping and Hybridized Random Forests. *Geoderma* **2021**, *385*, 114858. [[CrossRef](#)]

-
50. Taghizadeh-Mehrjardi, R.; Schmidt, K.; Toomanian, N.; Heung, B.; Behrens, T.; Mosavi, A.; Band, S.; Amirian-Chakan, A.; Fathabadi, A.; Scholten, T. Improving the Spatial Prediction of Soil Salinity in Arid Regions Using Wavelet Transformation and Support Vector Regression Models. *Geoderma* **2021**, *383*, 114793. [[CrossRef](#)]
 51. Ma, G.; Ding, J.; Han, L.; Zhang, Z.; Ran, S. Digital Mapping of Soil Salinization Based on Sentinel-1 and Sentinel-2 Data Combined with Machine Learning Algorithms. *Reg. Sustain.* **2021**, *2*, 177–188. [[CrossRef](#)]
 52. Kiss, J.; Bedard-Haughn, A. Predictive Mapping of Solute-rich Wetlands in the Canadian Prairie Pothole Region Through High-resolution Digital Elevation Model Analyses. *Wetlands* **2021**, *41*, 38. [[CrossRef](#)]
 53. Jenny, H. Factors of Soil Formation. *Soil Sci.* **1941**, *52*, 415. [[CrossRef](#)]
 54. Becker, W.R.; Ló, T.B.; Johann, J.A.; Mercante, E. Statistical Features for Land Use and Land Cover Classification in Google Earth Engine. *Remote Sens. Appl.* **2021**, *21*, 100459. [[CrossRef](#)]
 55. Gilabert, M.A.; González-Piqueras, J.; García-Haro, F.J.; Meliá, J. A Generalized Soil-Adjusted Vegetation Index. *Remote Sens. Environ.* **2002**, *82*, 303–310. [[CrossRef](#)]
 56. Miao, Y.; Mulla, D.J.; Robert, P.C. Identifying Important Factors Influencing Corn Yield and Grain Quality Variability Using Artificial Neural Networks. *Precis. Agric.* **2006**, *7*, 117–135. [[CrossRef](#)]
 57. He, Y.; Wei, Y.; DePauw, R.; Qian, B.; Lemke, R.; Singh, A.; Cuthbert, R.; McConkey, B.; Wang, H. Spring Wheat Yield in the Semiarid Canadian Prairies: Effects of Precipitation Timing and Soil Texture over Recent 30 Years. *Field Crops Res.* **2013**, *149*, 329–337. [[CrossRef](#)]