Improved parent material map disaggregation methods in the Saskatchewan prairies using historical bare soil composite imagery

Authors: Sorenson, P.T., Kiss, J., and Bedard-Haughn, A.K.

Source: Canadian Journal of Soil Science, 103(1) : 47-63

Published By: Canadian Science Publishing

URL: https://doi.org/10.1139/cjss-2021-0154

Canadian Journal of Soil Science

OPEN ACCESS | Article

Improved parent material map disaggregation methods in the Saskatchewan prairies using historical bare soil composite imagery

P.T. Sorenson, J. Kiss, and A.K. Bedard-Haughn

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

Corresponding author: **P.T. Sorenson** (email: preston.sorenson@usask.ca)

Abstract

The major drivers of soil variation in Saskatchewan at scales finer than the existing soil maps are parent material variance, slope position, and salinity. There is therefore a need to generate finer-scale parent material maps as part of updating soil maps in Saskatchewan. As spatially referenced soil point data are lacking in Saskatchewan, predictive soil mapping methods that disaggregate existing soil parent material maps are required. This study focused on investigating important environmental covariates to use in parent material disaggregation, particularly bare soil composite imagery (BSCI). Synthetic point observations were generated using an area-proportional approach based on existing soil survey polygons and a random forest model was trained with those synthetic observations to predict parent material classes. Including BSCI as environmental covariates increased model accuracy from 0.38 to 0.52 and the model Kappa score from 0.19 to 0.35 compared with models where it was not included. Models that included training points from all locations, regardless of whether BSCI was available, and included BSCI as environmental covariates had similar results to the BSCI model with an accuracy of 0.48 and a Kappa value of 0.30. Based on these results, BSCI is an important covariate for parent material disaggregation in the Saskatchewan Prairies. Future work to disaggregate soil classes based on slope position and salinity, and to combine those methods with parent material disaggregation is needed to generate detailed soil maps for the Canadian Prairies.

Key words: predictive soil mapping, bare soil composite, soil map dissaggregation, remote sensing, soil parent material

Résumé

Les principaux facteurs qui régissent la variation des sols en Saskatchewan à une échelle plus fine que celle des cartes pédologiques existantes sont la variance du matériau originel, la position de la pente et la salinité. Lorsqu'on actualisera les cartes du sol de la Saskatchewan, il faudrait donc produire des cartes plus précises du matériau originel. Puisqu'on manque de données ponctuelles avec références spatiales pour les sols de la province, il conviendrait d'élaborer des méthodes prévisionnelles qui dissocieront le matériau originel des cartes du sol actuelles. Les auteurs ont examiné d'importantes covariables environnementales qu'on pourrait utiliser à cette fin, notamment les images composites du sol nu. Pour cela, ils ont créé des observations ponctuelles synthétiques en recourant à une méthode de calcul proportionnel de la surface, articulée sur les polygones d'arpentage existants, et ont appliqué ces observations à un modèle à forêt aléatoire afin de prédire les classes de matériau originel. Ajouter les images composites du sol nu aux covariables environnementales fait passer l'exactitude du modèle de 0,38 à 0,52 et sa valeur Kappa de 0,19 à 0,35, comparativement à celles des modèles qui n'incluent pas de telles images. Les modèles formés avec des points de chaque endroit, que des images composites du sol nu soient disponibles ou pas, et comprenant de telles images dans leurs covariables environnementales donnent des résultats similaires à ceux du modèle avec les images composites du sol nu (exactitude de 0,48 et valeur Kappa de 0,30). De ces résultats on conclut que les images composites du sol nu constituent une importante covariable pour la dissociation du matériau originel dans les prairies de la Saskatchewan. Il faudrait entreprendre d'autres recherches afin de dissocier les classes de sol d'après la position de la pente et la salinité, et combiner ces méthodes à la dissociation selon le matériau originel en vue de produire des cartes pédologiques détaillées des Prairies canadiennes. [Traduit par la Rédaction]

Mots-clés : cartes prévisionnelles du sol, images composites du sol nu, dissociation des cartes pédologiques, télédétection, matériau originel

Introduction

The Province of Saskatchewan had extensive surveying and soil mapping activities conducted between 1958 and 1998. The resulting soil survey maps were produced at scales of 1:100 000 or coarser and are available in an easily accessible digital platform (SKSIS Working Group 2018). While these maps were essential for agriculture management and land use planning in Saskatchewan throughout the late 20th century, there is increasing demand for more detailed soil maps to support targeted land use planning, soil carbon management, and precision agriculture. To meet these needs, predictive soil mapping using machine learning tools, in combination with a suite of environmental covariates, has been an increasing focus of research in Saskatchewan and around the world (McBratney et al. 2003).

As there is currently limited public spatially explicit point data for predictive soil mapping in Saskatchewan, there is a need for mapping approaches that do not rely solely on field-collected point data for model training. Given the extensive, coarse-scale, soil maps available, there is the potential to disaggregate those maps using established techniques to generate improved soil maps across the agricultural region of Saskatchewan. Polygon disaggregation by generating synthetic training points based on polygon labels is one such approach (Holmes et al. 2015; Chaney et al. 2016). Initial polygon disaggregation approaches focused on assigning equal points to polygons and randomly assigning class labels based on their relative proportions within a polygon (Odgers et al. 2014). Improvements have since been made to polygon disaggregation by incorporating area-proportional sampling and informing class assignment based on soil–landscape relationships (Møller et al. 2019). Additionally, the use of random forest models with a single sampling procedure was found to be much more computationally efficient than approaches that used multiple sampling procedures and C5.0 decision trees, with only a slight decrease in predictive accuracy (Møller et al. 2019).

Parent material is one of the key soil-forming factors (Jenny 1941) and a major control of important soil properties in the Canadian Prairies. Currently, in Saskatchewan, the finest-scale parent material maps are the classifications included in Saskatchewan's detailed soil survey. More detailed parent material maps are necessary to map soil variation at finer scales. A previous study in Saskatchewan identified that finer-scale mapping of hydropedological parameters can delineate soil variation associated with topography, but this typically requires high-resolution digital elevation models (DEMs), which are not widely available in Saskatchewan (Pennock et al. 2014; Kiss and Bedard-Haughn 2021). Successful disaggregation of parent material maps may not depend on such data requirements, making it a potential approach to improve soil maps across the Saskatchewan prairies. There have been successful studies in disaggregating existing parent material maps in British Columbia (Heung et al. 2014; Bulmer et al. 2016), which did not rely on high-resolution DEMs for model environmental covariates.

One potential approach to improve soil parent material map disaggregation in the Saskatchewan Prairies is to incorporate bare soil composite imagery (BSCI) as an environmental covariate. BSCI is generated by mining archival remote sensing imagery to isolate pixels where bare soil is present and using the resulting data stack to generate a composite image of bare soil conditions. The creation of BSCI has become a possibility with the development of large-scale cloud computing infrastructure (Gorelick et al. 2017). This is particularly necessary in the Canadian Prairies as pixels with vegetation or dead plant residue cover that exceed 20% make direct estimation of soil properties difficult (Bartholomeus et al. 2007, 2008). Vegetation or residue-free images are rare in Saskatchewan in recent decades due to the widespread adoption of conservation tillage (Statistics Canada 2017). BSCI has been used in a variety of locations such as Germany (Rogge et al. 2018), Brazil (Safanelli et al. 2020), and globally (Demattê et al. 2020) for predicting soil properties such as soil organic carbon. Recently, BSCI has been used to successfully generate historical soil carbon maps and clay content maps for Saskatchewan (Sorenson et al. 2021). Both soil organic carbon and clay content vary with parent material, along with other factors, and therefore there is potential that BSCI could support parent material map disaggregation in Saskatchewan.

This study had two main objectives. The first objective was to evaluate the potential of disaggregation approaches to generate finer-scale parent material maps in Saskatchewan, specifically by using a more computationally efficient random forest model approach with a single set of synthetically generated training data derived from existing soil survey maps of soil parent materials. The second objective of this study was to investigate the value of BSCI for improving parent material disaggregation in Saskatchewan.

Materials and methods

Soil parent material context

Nearly the entire extent of Saskatchewan's soils is formed on unconsolidated transported parent materials, predominantly the result of glacial processes (Anderson and Cerkowniak 2010). The most common soil parent materials in Saskatchewan are glacial tills (often enriched with limestone-rich bedrock), followed by glaciolacustrine and fluvial deposits (Table 1). Due to the overall slope of the land and drainage to the northeast, large glacial lakes formed that caused extensive glaciolacustrine deposits associated with proglacial lakes (Anderson and Cerkowniak 2010). Additionally, high-energy streams from the melting glaciers were responsible for the creation of fluvial deposits along deltas that formed where streams entered glacial lakes. Minor areas of eolian landscapes are also present, which formed due to post-glacial wind erosion. Minor areas of bedrock, eolian, peat, and colluvium are also present. The existing parent material map, from Saskatchewan's detailed soil survey polygons, is provided in Fig. 1.

Remote sensing data

Remote sensing data for the agricultural regions of Saskatchewan (Fig. 1) were acquired using Google Earth Engine (Gorelick et al. 2017) to be used as environmental

Table 1. Distribution of parent material types in the detailed soil survey polygons of Saskatchewan (SKSIS Working Group 2018) and the National Pedon Database parent material class counts (Agriculture and Agri-Food Canada 2016).

	Soil survey polygons		National Pedon Database	
Parent material	Total area (ha)	Percentage (%)	Count	Percentage (%)
Not applicable	917 948	2.5	NA	NA
Bedrock	545 856	1.5	7	0.8
Colluvium	232 922	0.6	2	0.2
Eolian	385 221	1.0	9	1.0
Fluvial	7 687 880	20.8	365	39.1
Lacustrine	6 451 544	17.5	274	29.3
Lacustrine over till	3 458 704	9.4	130	13.9
Peat	455 332	1.2	NA	NA
Till	16 807 393	45.5	147	15.7
Total	36 942 800	100.0	934	100.0

covariates in the parent material disaggregation models. Landsat 7 Tier 1 surface reflectance data were acquired from 1999 to 2021, with several band indices calculated (Table 2). The median anthocyanin reflectance index (ARI), which is associated with anthocyanin pigments in plant foliage (Gitelson et al. 2001), was calculated for data from the months of July and August. This index was included as it has been useful for discriminating vegetation differences in Alberta for peatland mapping (DeLancey et al. 2019). The median canopy response salinity index (CRSI) was also calculated using data from the months of July and August, as it has been used to distinguish soil differences associated with salinity (Scudiero et al. 2015). The median normalized difference vegetation index (NDVI) was calculated for the months of July and October, separately. These months were selected as July corresponds to peak photosynthetic activity, and only non-arable cropland would have any photosynthetic activity in October. Non-arable cropland would be expected to be preferentially located on sandier parent materials or those with steeper slopes. The standard deviation of NDVI using data from May to October was also determined. Level 1 C data from Sentinel 2 were acquired for the months of May to October from 2015 to 2020 to generate the median red-edge inflection index (REIP). The REIP was included in the study, as it has been an important parameter for discriminating vegetation differences in Alberta (DeLancey et al. 2019). Each parameter was median focal filtered using a 3 × 3 window size and exported from Google Earth Engine at 30 m spatial resolution. The Google Earth Engine scripts for generating these covariates are available on GitHub (Sorenson 2021*a*).

Data from the Sentinel 1 platform were also retrieved from Google Earth Engine. Processing of the Sentinel 1 data was done using the methodology in Hird et al. (2017). The median vertical–vertical (VV) and vertical–horizontal (VH) backscatter was determined using data from May to October from 2015 to 2020. To focus more on broader landscape trends, rather than fine-scale variation, the data were median focal filtered with a 9 × 9 window size and exported at a 30 m spatial resolution, rather than the default 10 m, to match the resolution of the Landsat 7 data. The Google Earth Engine script for generating these covariates is available on GitHub (Sorenson 2021*a*).

Bare soil composite imagery

BSCI was generated using Google Earth Engine and the methodology from Sorenson et al. (2021). Atmospherically corrected Tier 1 surface reflectance data for Landsat 5 were retrieved for the months of July and August from 1984 to 1995. Landsat 5 data from this time period, rather than more recent Sentinel-2 or Landsat 7 and 8 data, were used because the total number of acres in Saskatchewan under conventional tillage peaked in 1970 and declined to only 18% of farmland by 2006 (Statistics Canada 2015). Therefore, more recent imagery is heavily affected by crop residues. The months were limited to July and August to ensure that crop residues from previous years had time to decompose, thereby reducing the influence of crop residues on BSCI. Clouds, shadows, and low-quality pixels were filtered using the Landsat quality assessment band (pixel_qa). The Landsat 5 data had seven bands: Band 1 corresponding to blue light (450–520 nm), Band 2 corresponding to green light (520–600 nm), Band 3 corresponding to red light (620–690 nm), Band 5 (1550–1750 nm), Band 6 (10 400–12 500 nm), and Band 7 (2080–2350 nm). Band 6 was not used in the analysis.

Image collections were filtered using NDVI, normalized burn ratio (NBR_2), normalized difference water index (NDWI) (Demattê et al. 2018), and normalized difference index 7 (NDI7) (Sorenson et al. 2021). Pixels were kept as likely representing the bare soil surface if the NDVI value was less than 0.3 (Demattê et al. 2020), NBR_2 was less than 0.1 (Demattê et al. 2020), NDWI was less than 0.5 (Du et al. 2016), and NDI7 was less than 0 (Kokaly et al. 2017). Additionally, any region mapped as pasture or grassland in Agriculture and Agri-Food Canada's 2019 Annual Crop Inventory was masked (Agriculture and Agri-Food Canada 2019). While 2019 does not match the years of the bare soil imagery collections, both land uses tend to be managed for long term. Following filtering and masking, the median reflectance values for each pixel were calculated, and then spatially filtered using a circular 10 × 10 median focal filter. The focal filtering was done to account for the 300 m location uncertainty reported in the

Fig. 1. Overview maps of imagery and parent material in Saskatchewan. (*a*) True colour red-green-blue median surface reflectance image (Landsat 7 Bands 3–2-1) for Saskatchewan from July and August from 1999 to 2021. The black points indicate the location of the National Pedon Database sample locations used for validation. The squares with yellow diagonal lines indicate the locations of the finer spatial resolution example maps. (*b*) Existing Saskatchewan soil parent material map. Coordinates are in UTM Zone 13 N NAD83.

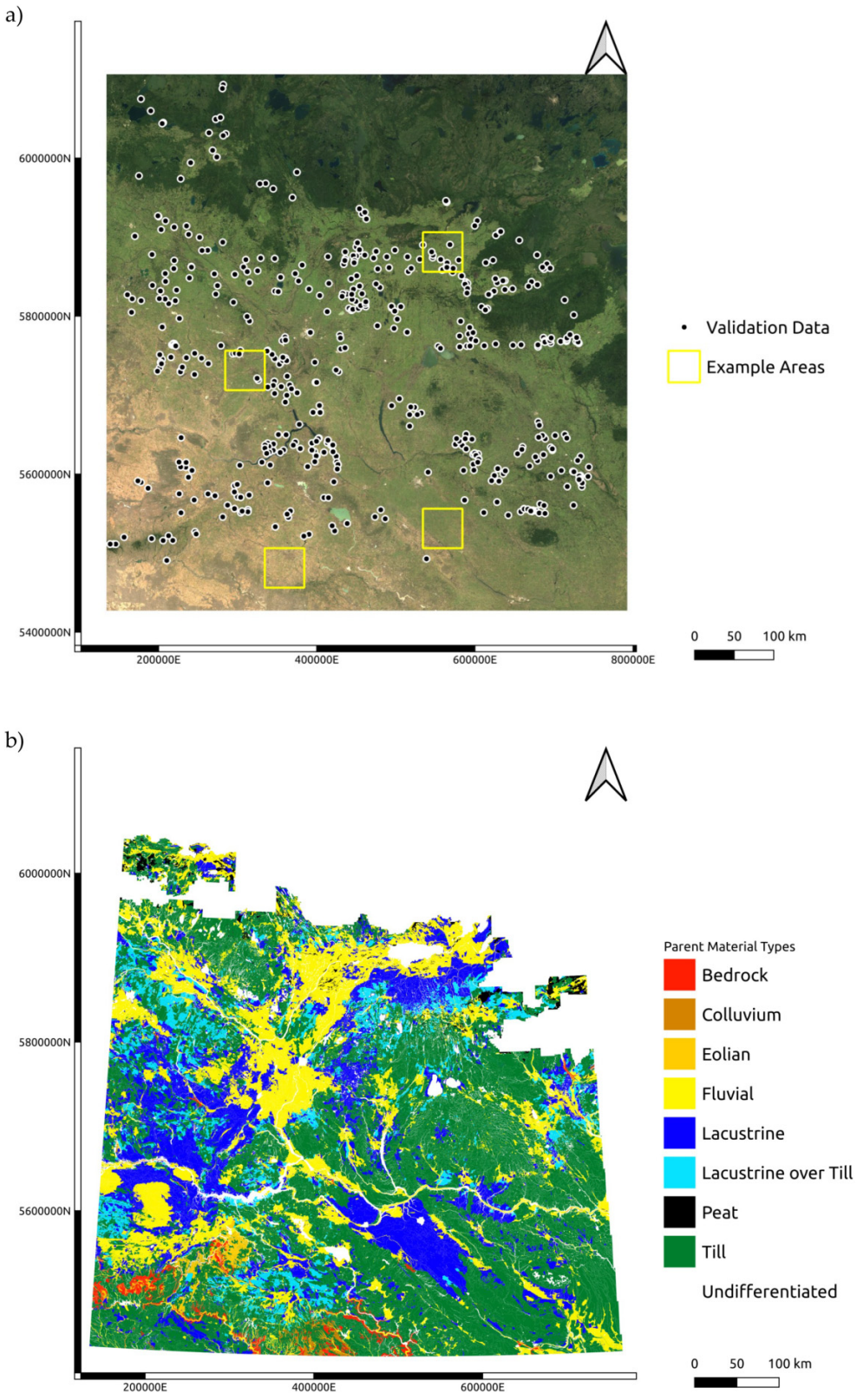


Table 2. Complete list of features included in the analysis prior to feature selection.

Feature	Data source	Band ratio equation
Band ratios		
Median anthocyanin reflectance index (ARI)	July and August Landsat 7 from 1999 to 2021	$\frac{\text{Band 4}}{\text{Band 1}} - \frac{\text{Band 4}}{\text{Band 2}}$
Median canopy response salinity index (CRSI)	July and August Landsat 7 from 1999 to 2021	$\left[\frac{(\text{Band4}\times\text{Band3})-(\text{Band2}\times\text{Band1})}{(\text{Band4}\times\text{Band3})+(\text{Band2}\times\text{Band1})}\right]^{1/2}$
Median normalized difference vegetation index (NDVI)	July Landsat 7 from 1999 to 2021, and October Landsat 7 from 1999 to 2021	$\frac{\text{Band 4}-\text{Band 3}}{\text{Band 4}+\text{Band 3}}$
Standard deviation of NDVI	May to October Landsat 7 from 1999 to 2021	
Median red-edge inflection index (REIP)	May to October Sentinel 2 from 2015 to 2020	$702 + 40\left[\frac{(\text{Band 4}+\text{Band 7})/2 - \text{Band 5}}{\text{Band 6}- \text{Band 5}}\right]$
Feature	**Data source**	**Filtering levels**
Terrain attributes		
Terrain ruggedness index (TRI) 10 × 10 window size	Advanced land observation satellite (ALOS) digital surface model	• No spatial filtering • 3 × 3, 5 × 5, and 9 × 9 median focal filtering of the input surface model • 3 × 3 median focal filtering of the output data
TRI 20 × 20 window size	ALOS digital surface model	• 9 × 9 median focal filtering of the input surface model
Standard deviation of elevation	ALOS digital surface model	• 3 × 3 focal window with 3 × 3 median focal filter of the input surface model • 5 × 5 focal window with 3 × 3 median focal filter of the input surface model • 9 × 9 focal window with 3 × 3 median focal filter of the input surface model • 21 × 21 focal window with 3 × 3 median focal filter of the input surface model • 21 × 21 focal window with 9 × 9 median focal filter of the input surface model • 101 × 101 focal window with 9 × 9 median focal filter of the input surface model
Feature	**Data source**	**Bands**
Remote sensing bands		
Bare soil composite imagery (BSCI)	July and August Landsat 5 from 1984 to 1995	Band 1, Band 2, Band 3, Band 4, Band 5, and Band 7
Sentinel 1 synthetic aperture radar	May to October Sentinel 1 with 9 × 9 median focal filtering from 2015 to 2020	Vertical–vertical (VV) and Vertical–horizontal (VH) polarization.

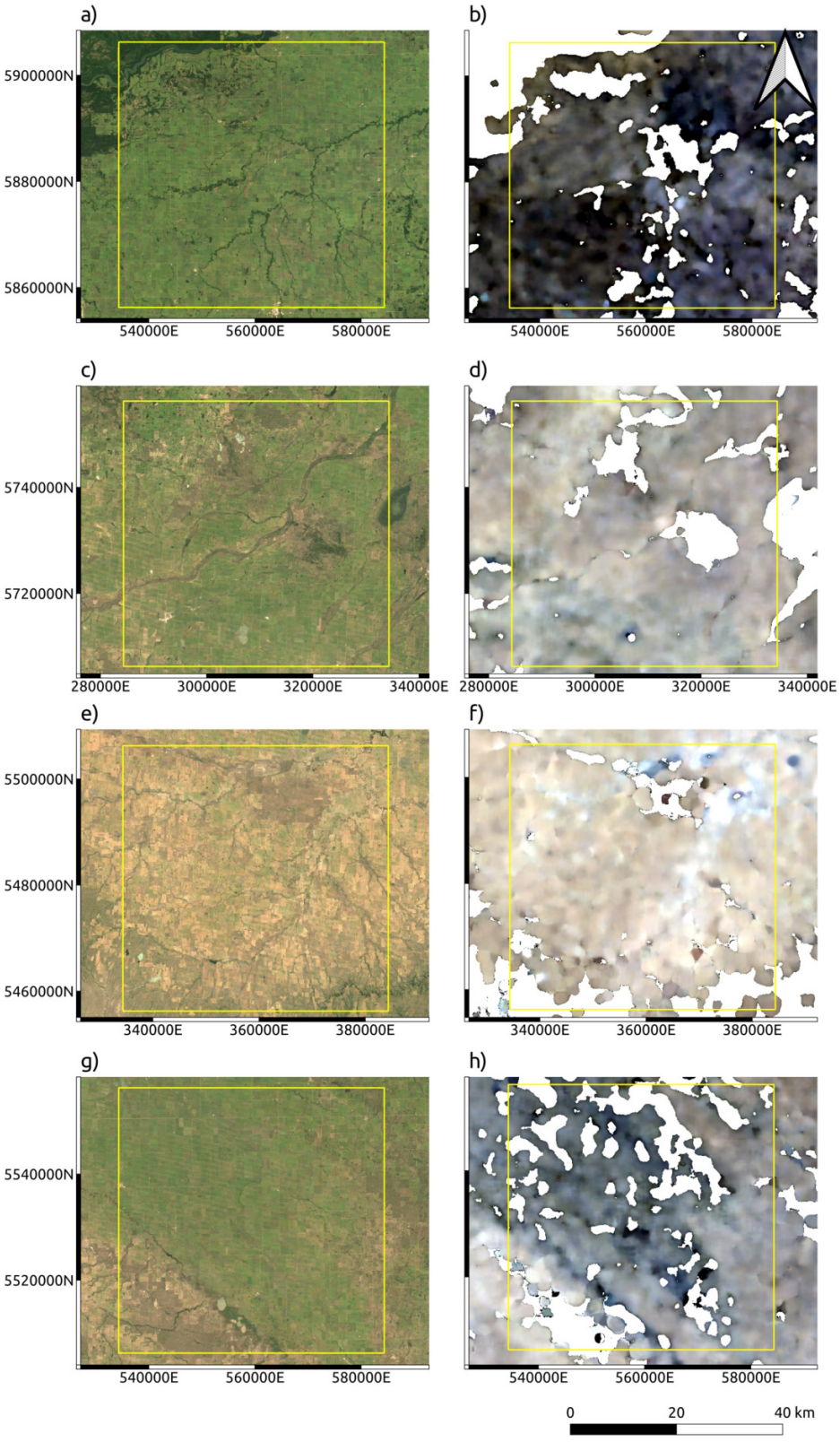
Note: Band ratio equations for each band ratio were used as an environmental covariate in the analysis. The listed bands correspond to Landsat 5 and 7 except for the red-edge inflection point which correspond to Sentinel 2 bands.

National Pedon Database and to smooth out fine-scale variation more likely attributable to factors other than parent material. The final filtered composite image was then aggregated to a spatial resolution of 30 m. Additional details and band ratio equations are available in Sorenson et al. (2021) and the Google Earth Engine script is available on GitHub (Sorenson 2021*b*). Example results for the BSCI are provided in Fig. 2, which includes true colour red-green-blue images of the bare soil surface.

Terrain attributes

Terrain attributes were determined using the 30 m digital surface model from the Advanced Land Observation Satellite (ALOS) (Japan Aerospace Exploration Agency 2015) retrieved using Google Earth Engine. Two terrain attributes, terrain ruggedness index (TRI) and the standard deviation of elevation, were calculated using a range of median focal filtering window sizes and calculation window sizes (Table 2). Median focal filtering was used, rather than mean, as it is less sensitive to the effects of outliers. These attributes were selected, as they represent coarser scale patterns in landscape variability rather than specific fine-scale patterns related to factors such as local landscape morphometry, hydrological characteristics, and landscape exposure. While these finer-scale attributes were documented to improve predictive parent material mapping at 100 m in British Columbia

Fig. 2. Imagery for the four finer spatial resolution example areas. The first row corresponds to the northeast example area, the second row is the northwest example area, the third row is the southwest example area, and the fourth row is the southeast example area. The figures in column (*a*) are true colour red-green-blue median surface reflectance image (Landsat 7 Bands 3-2-1) for Saskatchewan from July and August from 1999 to 2021. The figures in column (*b*) are true colour red-green-blue median reflectance bare soil pixel composite images (Landsat 5 Bands 3-2-1) for Saskatchewan from 1984 to 1995. This represents a true colour image of the exposed soil surface. The white areas in column (*b*) correspond to areas either without bare soil pixels present or permanent pasture areas that have been masked. Coordinates are in UTM Zone 13 N NAD83.



(Heung et al. 2014; Bulmer et al. 2016), the characteristics of prairie landscapes make these attributes challenging to accurately calculate with currently available DEMs in Saskatchewan. Satellite-derived DEMs with spatial resolutions of 30 m, like the ALOS model used here, lack the horizontal and vertical accuracy to capture fine-scale variation in the relatively low-relief Canadian Prairies. Specifically, the ALOS 30 m DEM has a reported vertical root mean square error of 1.78 m (Caglar et al. 2018), which is greater than the relief differences driving soil property differences in some hummocky landscapes in Saskatchewan (Landi et al. 2004). Additionally, 30 m spatial resolution is not fine enough to consistently characterize slope and topographic variation in many Saskatchewan landscapes (Landi et al. 2004). However, currently these are the best freely available DEM data for Saskatchewan. Therefore, this study focused on using terrain attributes that accounted for general degrees of variability more broadly.

TRI was calculated using the system for automated geoscientific analyses (SAGA) (Conrad et al. 2015). The TRI was calculated with either a window size of 10 or 20 (Table 2), and with the following median focal filters applied to the DEM prior to processing: no filter, 3 × 3, 5 × 5, and 9 × 9 (Table 2). Applying a 3 × 3 median focal filter to the TRI results with no filter on the DEM was also tested. The standard deviation of elevation was calculated using 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 (Table 2). The standard deviation of elevation was also calculated using window sizes of 21 × 21 and 101 × 101 for a 9 × 9 median focal-filtered DEM. Examples of the terrain features determined as important environmental covariates during model development are illustrated in Fig. 3. TRI was not calculated using a 101 × 101 window because of the computational time required compared to standard deviation. Multiple scales were considered as multi-scale digital terrain analysis has been shown to improve predictive soil mapping results (Behrens et al. 2010).

Model development

Predictive models were developed using the detailed soil survey polygons as training data, which were mapped at a 1:100 000 scale. The disaggregation approach was similar to Møller et al. (2019) where a single set of area-proportionally sampled synthetic training points were generated and a random forest model was trained. This approach was selected for the computational efficiency (Møller et al. 2019). Synthetic training data were developed by randomly assigning points within each polygon on an area proportional basis. The number of points generated for each polygon was equal to the natural logarithm of the polygon area divided by the area of the smallest polygon in the data set. This number was then multiplied by 5 for the final number of points per polygon. The rationale for this approach was that it ensured multiple training points are generated for each polygon and that larger polygons have more points generated to reduce the likelihood that they are underrepresented in the training data.

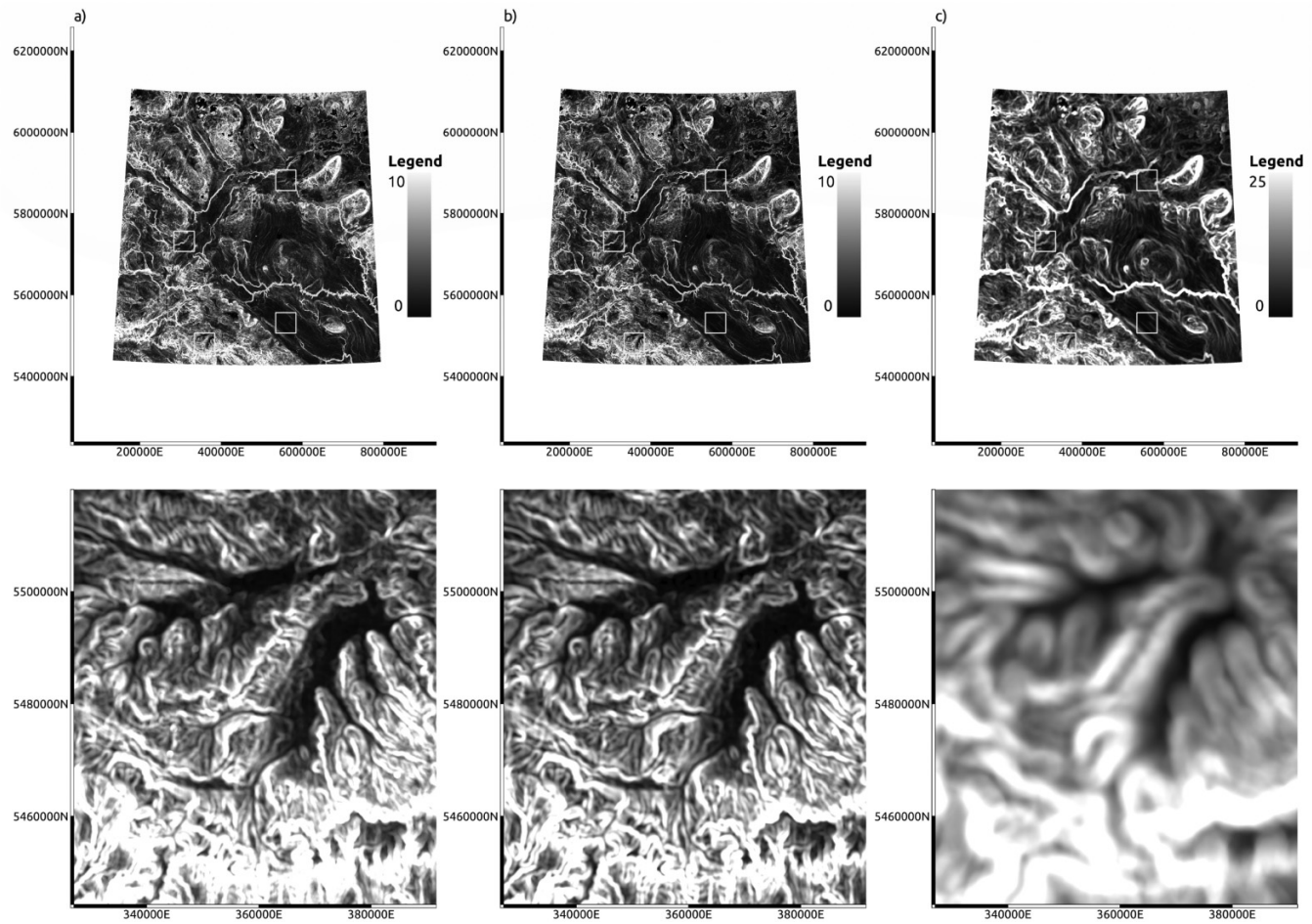
Points were assigned to the parent material class based on the first deposition layer described for the polygon that they were generated in, with the exception of the *lacustrine over till* class, which reflects both the first and second deposition layers. Polygons with parent material values of *undifferentiated* indicated that there is too much variability in parent materials at the mapping scale of the detailed soil survey polygons to assign a parent material class, so data from these polygons were not used for model development. As each polygon has a single component, the single component was used to assign the training data label. Multi-component polygons in the parent material maps were labelled as *undifferentiated* and they were excluded during training data creation. In total, 5 786 772 synthetic training points were generated. The scripts for generating the training data from the parent material polygons are available on GitHub (Sorenson 2021*c*).

The parent material classes for the synthetic training points generated from the detailed soil survey polygons were simplified. *Fluvial eolian* and *fluvial lacustrine* parent materials were grouped as *fluvial* because of the similar characteristics of these materials. Additionally, minor parent materials with very similar characteristics were removed, specifically *glaciolacustrine* and *glaciofluvial*. *Fluvial eolian* and *fluvial lacustrine* were grouped as *fluvial*. *Fen* and *sphagnum peat* were grouped as a single *peat* class. *Lacustro-till* were relabelled as *lacustrine* due to its infrequent occurrence and high similarity to *lacustrine* deposits. *Residual* and *undifferentiated bedrock* were combined into a *bedrock* class. The final parent material classes were: *bedrock, colluvium, eolian, fluvial, lacustrine, lacustrine over till, peat*, and *till*.

To improve the computational efficiency of the model building, the total points were randomly subset to three sets with 1 000 000 training points from the original 5 786 772. Separate training subsamples were generated for the three following models: (1) the *BSCI-only model*, which includes only training points from locations where BSCI is available and includes BSCI as environmental covariates, (2) the *exhaustive model*, which includes training points from all locations, regardless if BSCI is available and includes BSCI as environmental covariates, and (3) the *BSCI-excluded model*, which also includes training points from all locations but does not include BSCI as environmental covariates. Where BSCI imagery was not available, a null value of 0 was set for model development purposes.

Following creation of the training data, random forest models were built using the *ranger* package in R (Wright and Ziegler 2017). As class imbalances exist in the training data, the case weights term in the *ranger* model was used to balance across classes during model building. The case weights term in the *ranger* package allows for probabilities to be assigned to each training point that determines the likelihood a point will be sampled during individual tree building in the random forest. As a result, each tree of the random forest can be built with balanced classes, and the overall random forest model will therefore be built on class balanced data even if the original data set has imbalanced classes. Case weights were set so that each parent material class had an equal chance of being subsampled for each individual tree

Fig. 3. Terrain attribute imagery for the three terrain parameters selected as important predictors by the feature selection process. The first row provides imagery for the portion of Saskatchewan with conventional soil survey maps at 1:100 000. The bottom row is the imagery for the southwest example area. The figures in column (*a*) are the ALOS DEM with a 3 × 3 median filter applied and the standard deviation of elevation calculated with a 21 × 21 window. Figs. in column (*b*) are the ALOS DEM with a 9 × 9 median filter applied and the standard deviation of elevation calculated with a 21 × 21 window. The figures in column (*c*) are the ALOS DEM with a 9 × 9 median filter applied and the standard deviation of elevation calculated with a 101 × 101 window. Coordinates are in UTM Zone 13 N NAD83.



built in the random forest. An initial model was built using all environmental covariates listed in Table 2 to determine feature importance based on Gini index. Features were then selected for the final models using a backward feature selection process, where random forest models were built using sequentially less features. For each iteration, the feature importances were recalculated, and the least important feature was removed. The final features selected were determined based on where the out-of-bag classification error was minimized. In total, the *BSCI-only model* used 16 features, the *exhaustive model* used 20 features, and the *BSCI-excluded model* used 11 features (Table 3). Final random forest models were built for each scenario with the following model hyperparameters: probability set to true, extra trees as the split rule, and case weights set so that subsampling probability was equalized across parent material classes to reduce the influence of class imbalances on the model. The scripts for generating the random forest models and resulting parent material maps are available on GitHub (Sorenson 2021*c*).

Validation data

Data from the National Pedon Database (Table 1) were used as model validation data (Agriculture and Agri-Food Canada 2016). Historically, the samples were collected to represent modal soils from dominant soil types from a region and were collected by horizon with analyses performed on the homogenized horizon samples. The location accuracy for these samples is low, with specific locations estimated relative to land marks by the soil surveyors. The estimated location uncertainty is approximately ±300 m. The 300 m uncertainty is the location uncertainty reported for the data in the National Pedon Database. These soil pedon observations were used to inform the soil survey parent material mapping and so they are not completely independent from the maps used as training data, but the point observations themselves were not used in the model training. The soil parent material classes were simplified for this study based on the original parent material classes in the survey database using the same approach as the training data.

Table 3. Final features used for each of the three predictive models along with feature importance values for the final model.

Feature	Importance
BSCI-only model	
Sentinel 1 VH polarization	61 115
Standard deviation of 9 × 9 median filtered elevation with a 101 × 101 focal window	51 431
Standard deviation of NDVI	50 107
Median ARI	49 904
Bare Soil Composite Band 5	48 257
Bare Soil Composite Band 1	44 765
Median October NDVI	43 946
Bare Soil Composite Band 7	43 831
Bare Soil Composite Band 3	43 756
Bare Soil Composite Band 2	43 657
Sentinel 1 VV polarization	43 423
Bare Soil Composite Band 4	43 141
Standard deviation of 3 × 3 median filtered elevation with a 21 × 21 focal window	40 012
Standard deviation of 3 × 3 median filtered elevation with a 9 × 9 focal window	36 929
Bare Soil Composite Band 6	35 721
Median July NDVI	35 564
Exhaustive model	
Sentinel 1 VH polarization	55 924
Standard deviation of 9 × 9 median filtered elevation with a 101 × 101 focal window	48 665
Sentinel 1 VV polarization	47 296
Median ARI	42 597
Standard deviation of NDVI	41 207
Median REIP	35 934
Median October NDVI	35 181
Median July NDVI	33 683
Standard deviation of 3 × 3 median filtered elevation with a 21 × 21 focal window	33 393
Standard deviation of 9 × 9 median filtered elevation with a 21 × 21 focal window	32 690
Bare Soil Composite Band 5	32 644
Bare Soil Composite Band 7	32 160
Median CRSI	31 071
TRI 20 × 20 window size with 9 × 9 median focal filtering	30 248
Standard deviation of 3 × 3 median filtered elevation with a 9 × 9 focal window	29 528
Bare Soil Composite Band 6	28 869
Bare Soil Composite Band 1	28 533
TRI 10 × 10 window size with 3 × 3 median focal filtering	27 851
Bare Soil Composite Band 2	27 420
Bare Soil Composite Band 4	26 981
BSCI-excluded model	
Sentinel 1 VH polarization	88 441
Standard deviation of 9 × 9 median filtered elevation with a 101 × 101 focal window	70 274
Sentinel 1 VV polarization	68 883
Standard deviation of NDVI	64 801
Median ARI	63 417
Median October NDVI	61 536
Median REIP	57 571
Standard deviation of 3 × 3 median filtered elevation with a 21 × 21 focal window	55 560
Median July NDVI	53 353
Standard deviation of 9 × 9 median filtered elevation with a 21 × 21 focal window	52 061
Median CRSI	50 442

Note: These models include only training data with BSCI present, using training data with and without BSCI, and without using BSCI.

Table 4. Confusion matrix for the independent validation results of the *BSCI-only model*.

Most likely parent material prediction								
	Predicted class							
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Class producer accuracy
Bedrock	1		1	1		1	1	0.2
Colluvium							1	0
Eolian			6				2	0.75
Fluvial			4	149	30	25	25	0.64
Lacustrine				42	114	39	46	0.47
Lacustrine over till				18	49	29	18	0.25
Till	1			8	16	17	76	0.64
Overall accuracy	0.52							
Kappa	0.35							
Second most likely parent material prediction								
	Predicted class							
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Class producer accuracy
Bedrock	1		1				3	0.2
Colluvium	1							0
Eolian	1		1	2		1	3	0.13
Fluvial			4	32	53	43	101	0.14
Lacustrine				40	50	64	87	0.21
Lacustrine over till				14	18	37	45	0.32
Till			3	25	37	34	19	0.16
Overall accuracy	0.19							
Kappa	−0.04							

The distribution of parent material classes amongst pedon validation data did not match the distribution of parent material classes reflected by the detailed soil survey polygons (Table 1). This was likely because the pedon data were collected to provide example profiles for the range of soil associations present in Saskatchewan, not to sample parent material classes on a proportional basis. The most frequent parent material in the pedon validation data was *fluvial*, followed by *lacustrine*, *till*, and then *lacustrine over till* (Table 1). Minor amounts of *eolian*, *bedrock*, and *colluvium* parent materials were present. No *peat* data are present in the validation data. For the detailed soil survey polygon data, the most common parent material class was *till*, followed by *fluvial* and then *lacustrine* (Table 1).

Model validation

Model performance was evaluated based on overall model accuracy, Cohen's Kappa, which measures inter-rater reliability and considers agreement occurring by chance, and specific class producer accuracy. Accuracy and Kappa values in this study are reported on a scale of 0–1. Producer accuracy refers to the number of correctly predicted observations of a soil class per the total number of observations of that class (Malone et al. 2017), and is referred in the results simply as class accuracy.

Uncertainty estimates

Prediction confidence was determined using the confusion index (Burrough et al. 1997; Odgers et al. 2011; Brungard et al. 2015), which is calculated with the following equation:

$$CI = \{1 - [u_{\max} - u_{(\max-1)}]\}$$

where CI is the confusion index, $u_{\max}$ is the probability value of the most likely prediction based on the random forest proportion of votes in the ensemble, and $u_{(\max-1)}$ is the probability value for the second most likely prediction based on the proportion votes in the ensemble. Confidence interval values closer to 1 indicate less certainty and values closer to 0 indicate more certainty in the prediction.

Results and discussion

The results of this study are presented in two sections. The first section is a review of the results for the *BCSI-only model*, the *BCSI-excluded model*, and the *exhaustive model*. The second section provides details on the model features

Table 5. Confusion matrix for the independent validation results of the *exhaustive model*.

Most likely parent material prediction									
	Predicted class								
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Peat	Class producer accuracy
Bedrock			2	1		1	1		0
Colluvium	1								0
Eolian			5				3		0.63
Fluvial			8	118	38	32	33	4	0.52
Lacustrine				20	122	54	45		0.55
Lacustrine over till				11	50	36	17		0.52
Till	1			7	19	29	62		0.53
Overall accuracy	0.48								
Kappa	0.30								

Second most likely parent material prediction									
	Predicted class								
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Peat	Class producer accuracy
Bedrock	1						4		0.2
Colluvium							1		0
Eolian			1	2		2	3		0.13
Fluvial			4	41	64	46	78		0.18
Lacustrine	2			44	48	65	82		0.20
Lacustrine over till	1			11	22	40	40		0.10
Till			2	11	38	32	35		0.30
Overall accuracy	0.23								
Kappa	0.01								

that were most important amongst each of the three model types.

Model performance

Overall, the *BSCI-only model* was the best performing model. The accuracy for the most probable parent material was 0.52 and Kappa was 0.35 (Table 4). The model performance for the *exhaustive model* was very similar, but slightly lower, with an overall accuracy of 0.48 and Kappa of 0.30 (Table 5). The model performance for the *BSCI-excluded model* was lowest, with an accuracy of 0.38 and a Kappa value of 0.19 (Table 6). Predictive results were comparable between the two models that included BSCI environmental covariates (Fig. 4). However, it is important to note that the *BSCI-only model* could only generate predicted maps for those regions where BSCI imagery exists, which is a significant limitation (Fig. 5).

In comparison, classification accuracy for regional-scale parent material mapping in British Columbia had an overall accuracy value of 0.779 and Kappa values of 0.697 (Heung et al. 2014). That study focused on topographic derivatives that appear to work well in British Columbia. The greater relief in British Columbia improves results with coarser-scale DEMs with higher vertical error, compared to that in the Canadian Prairies where relief changes related to changes in soil properties can be shallower than the DEM vertical error (Landi et al. 2004). The better results in British Columbia compared to this Saskatchewan study might be explained by the less spatially extensive study area mapped in the British Columbia study (5472 km^2 in British Columbia versus 298 000 km^2 in Saskatchewan). A second British Columbia study targeting a more extensive study area (945 000 km^2) had comparable accuracies to this study with values of 0.411 or 0.415, depending on whether a balanced or constrained approach was used for predicting parent material (Bulmer et al. 2016)

Cross-class confusion for the *BSCI-only model* showed variance in performance across the parent material classes. Due to the lack of validation points, the mapping performance for *colluvium* and *peat* classes cannot be evaluated as part of this study (Table 4). *Eolian* and *bedrock* classes had few validation points resulting in less certainty in the overall predictions for those classes. The *bedrock* class observations were predicted across the parent material classes. Predictive accuracy for the *eolian* parent material class was the highest with a class accuracy of 0.75 (Table 4). *Fluvial* and *till* parent materials had the

Table 6. Confusion matrix for the independent validation results of the *BSCI-excluded model*.

Most likely parent material prediction									
	Predicted class								
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Peat	Class accuracy
Bedrock				2	1	2	1		0
Colluvium	1						1		0
Eolian	1		2		3		3		0.22
Fluvial	4		4	130	66	49	100	12	0.37
Lacustrine			1	42	116	64	48	3	0.43
Lacustrine over till			1	16	46	48	17	2	0.38
Till	3		4	17	27	27	62	6	0.44
Overall accuracy	0.38								
Kappa	0.19								

Second most likely parent material prediction									
	Predicted class								
Actual class	Bedrock	Colluvium	Eolian	Fluvial	Lacustrine	Lacustrine over till	Till	Peat	Class accuracy
Bedrock				1	1	2	2		0
Colluvium	1						1		0
Eolian		1	2	1	1	3	1		0.22
Fluvial	3			87	84	69	95	27	0.26
Lacustrine		2		63	77	57	73	2	0.28
Lacustrine over till		1		17	41	27	43	1	0.21
Till	1			31	43	21	46	4	0.32
Overall accuracy	0.26								
Kappa	0.03								

majority of the validation points classified correctly, with the primary misclassification for *fluvial* being *lacustrine* (Table 4). Interestingly, an area mapped as *till* in the southwest corner of the southeast example area was mapped as *fluvial* deposits rather than *till* even though it was mapped as *till* in the original detailed soil polygons (Fig. 4). Without validation data in this area, the accuracy of this mapping cannot be confirmed. However, some of this area could include tills reworked by fluvial processes as they are adjacent to a large glacial lake that was present in this region of Saskatchewan (SKSIS Working Group 2018).

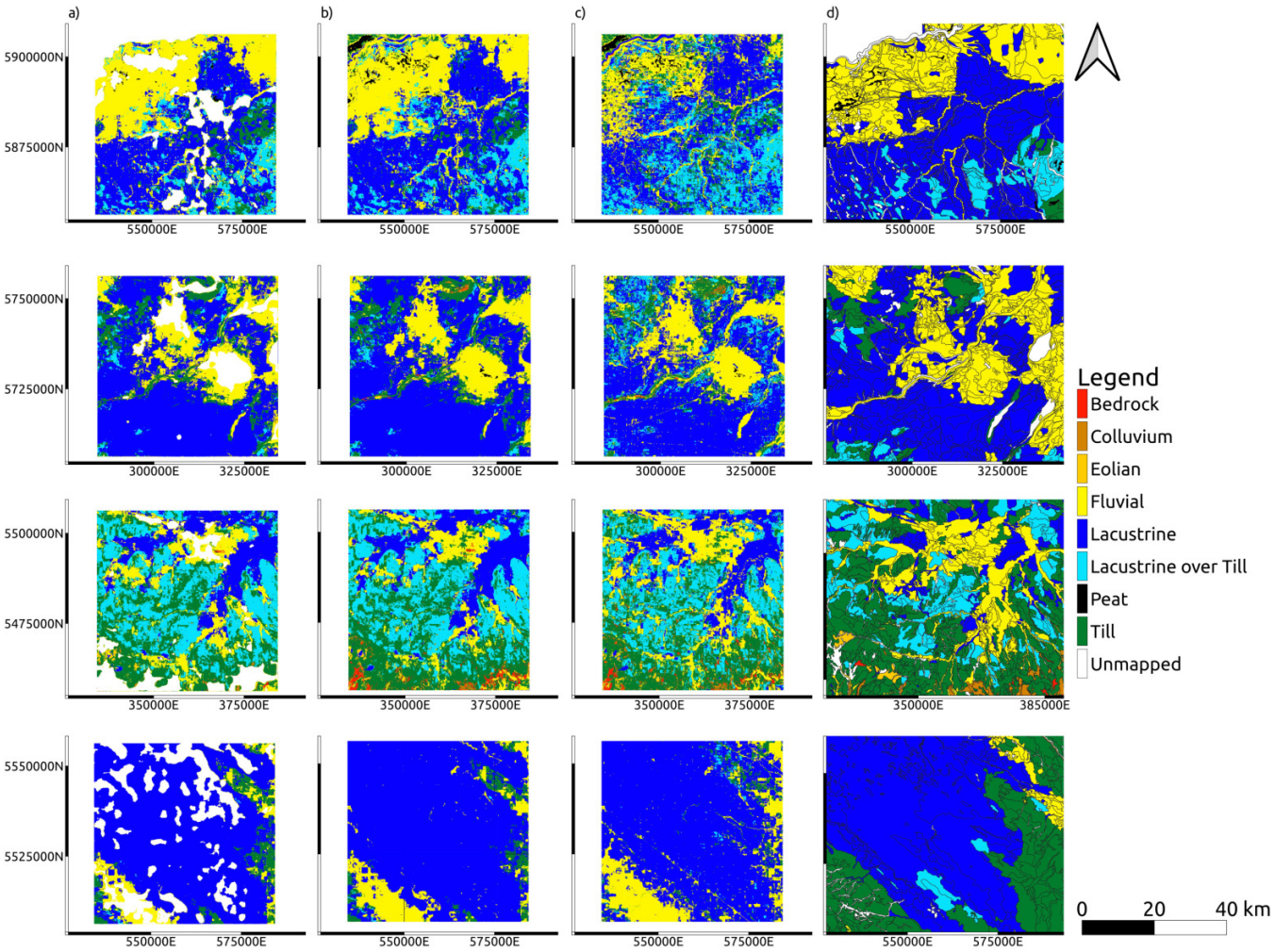
Till was primarily misclassified as *lacustrine over till*, which is the parent material with the most similar characteristics, and surface expression influenced by the underlying *Till*. Predictions for *lacustrine* parent material observations were less than 50% correct, with a class accuracy of 0.47. There was a fairly even distribution of misclassifications across *fluvial*, *lacustrine over till,* and *till* observations. *Lacustrine over till* observations were classified correctly 25% of the time, with the class being primarily misclassified as *lacustrine*. This is likely because of the importance of bands 5 and 7, which are related to clay content (Sorenson et al. 2021), to the model, and the similar surface textures between these two parent materials.

Cross-class confusion for the *exhaustive model* (Table 5) was similar to the *BSCI-only model*. Overall, this model had slightly lower class accuracies for the *eolian*, *fluvial*, and *till* parent material classes. There was a slight increase in class accuracies for the *lacustrine* and *lacustrine over till* classes. There were major changes in the performances per class for the *BSCI-excluded model*. The accuracies of all classes dropped (Table 6). This was particularly the case for the *eolian* class. This is likely because of the lack of bare soil imagery Bands 5 and 7 to identify the lower clay content areas associated with *eolian* parent materials. Based on these shifts in overall accuracy, Kappa, and individual class accuracies, the utility of the *BSCI-excluded model* is low.

Feature importance

Overall, the most important feature for the *BSCI-only model* was the Sentinel-1 VH backscatter (Table 3). A possible explanation for the importance of this feature is that VH backscatter has been shown to be related to aboveground biomass (Laurin et al. 2018). By comparison, the VV backscatter was less important (11th most important feature). VV is sensitive to soil moisture, but does vary with vegetation characteristics as well (Vreugdenhil et al. 2018). Particularly under

Fig. 4. Parent material prediction results for each of the four example areas illustrated in Fig. 1. The first row corresponds to the northeast example area, the second row is the northwest example area, the third row is the southwest example area, and the fourth row is the southeast example area. The figures in column (*a*) are the prediction results from the *BSCI-only model*. The figures in column (*b*) are the prediction results from the *exhaustive model*. Figs. in column (*c*) are the prediction results from the *BSCI-excluded model*. The figures in column (*d*) are the detailed soil survey polygons displayed by parent material. Coordinates are in UTM Zone 13 N NAD83.

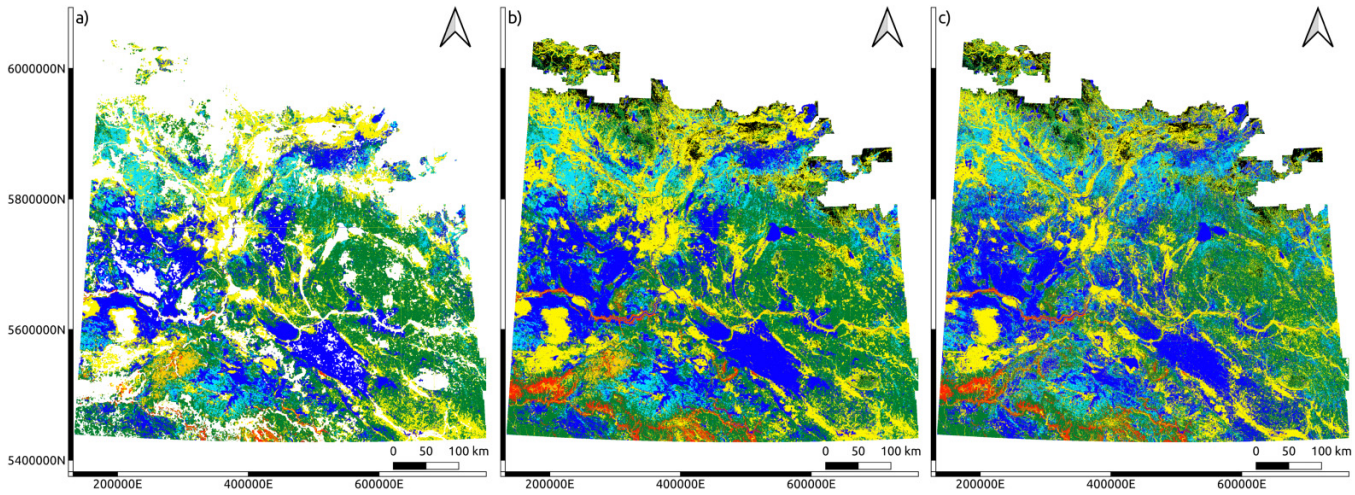


water-limiting conditions, finer textured soils have increased yields and associated biomass (He et al. 2013), and parent materials having higher average clay contents, and therefore biomass, are likely the reason Sentinel-1 VH backscatter was an important feature. Additionally, increased biomass would be associated with areas managed as pasture, which is also associated with coarser parent materials such as *fluvial* deposits.

The second most important feature was the standard deviation of 9 × 9 median focal filtered elevation with a 101 × 101 focal window. This terrain feature was the most smoothed and coarsest-scale feature of the terrain attributes (Fig. 3), suggesting that the coarser-scale terrain attributes were more useful for parent material mapping in Saskatchewan, at least given the ALOS DEM used in this study. The coarse smoothing resulted in a terrain feature associated with larger-scale landscape heterogeneity. This likely was helpful in differentiating between *glacial till* with heterogeneous hummocky landforms compared to level or undulating *lacustrine* landscapes that would have low amounts of terrain variability.

Optical remote sensing features were then the next most important features, with the median ARI being one of the most important of these features (Table 3). The ARI is a function of anthocyanin pigments in foliage (Gitelson et al. 2001) and is likely corresponding to variation in plant composition, which would be closely related to management practices. Coarser-textured hummocky parent materials, such as *fluvial* parent materials, are more likely to be managed as pasture compared to annually cropped *lacustrine* deposits. The standard deviation of NDVI was the most important NDVI feature (third-most important feature overall). Areas with higher NDVI standard deviations indicated more changes 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. The standard

Fig. 5. Parent material prediction results for the agricultural region of Saskatchewan. The figures in column (*a*) are the prediction results from the *BSCI-only model*, column (*b*) are the prediction results from the *exhaustive model*, and column (*c*) are the prediction results from the *BSCI-excluded model*. Coordinates are in UTM Zone 13 N NAD83.



deviation of NDVI has been shown to be a useful remote sensing feature for land use monitoring and land classification (Becker et al. 2021). The median October NDVI and Median July NDVI were less important as the eleventh and fourteenth most important features, respectively (Table 3).

While not the most important features, BSCI features were reasonably important: the fifth and eighth most important features were the BSCI bands 5 and 7 (Table 3), which correspond to shortwave infrared bands. The importance of these bands could be due to increased absorption in these wavelengths with higher clay contents considering they were found to be the most important bands for mapping clay content in Saskatchewan with BSCI (Sorenson et al. 2021). Clay has absorption features within Landsat 5's shortwave bands range (Rossel et al. 2010). However, these bands cover a wide portion of the shortwave infrared portion of the electromagnetic spectrum that may reflect other soil properties such as soil organic carbon, so their relationship to clay content is not certain. The sixth, ninth, and tenth most important features were the visible light bands for the BSCI. These bands were important predictors for SOC prediction in Saskatchewan (Sorenson et al. 2021), along with the near-infrared band which was the 12th most important feature. Silt and clay content, which are affected by parent material type, has been documented to be positively related to soil organic carbon concentration in Saskatchewan (Plante et al. 2006).

Standard deviation of a 3 × 3 median-filtered elevation with a 21 × 21 focal window was the 13th most important feature, which would be reflective of finer-scale terrain patterns compared to the 9 × 9 filtered 101 × 101 standard deviation (Fig. 3). Including a feature that characterizes the terrain patterns at different scales likely helped distinguish different parent materials as the contrasts amongst some parent materials were greater at coarser scales (i.e., *lacustrine* versus *lacustrine over till*) and others were greater at finer scales (i.e., *lacustrine* versus *fluvial*). While inclusion of finer-scale terrain attributes did improve the results, the coarsest-scale terrain attribute was more important (Table 3), at least for mapping parent material in Saskatchewan with a 30 m DEM, and in conjunction with other remote sensing features. While TRI has value quantifying topographic heterogeneity in other prediction soil mapping studies (Brungard et al. 2015), the simple standard deviation of elevation was determined to be more important by the random forest models in this study for the *BCSI-only* mapping.

The most important features for the *exhaustive model* were very similar to the *BSCI-only model* (Table 3). There were, however, differences in terms of the relative importance of each feature. The importance of Sentinel-1 VV backscatter was much greater for the *exhaustive model* than for the *BSCI-only model*. The likely reason for this is that because this model included areas where bare soil composite data were not present, the model needed to rely on other feature to distinguish differences amongst parent material classes. There was also an additional terrain feature included in the final selected feature set, particularly different focal window sizes for the standard deviation of elevation and two TRI features. In contrast with the *BSCI-only model*, BSCI band 3 was not included in the final model. Overall, BSCI was less important and bands 5 and 7 were found to be the most important BSCI bands. The reduction in importance of the BSCI bands could be partially explained by the fact that a significant portion of the training data would not have variance in bare soil values.

The median REIP and CRSI were also important features for the *exhaustive* model. The REIP is an approximation of the near-infrared red inflection point in vegetation spectra. The location of REIP is highly correlated with foliar chlorophyll content, photosynthetic activity, and a good predictor of leaf area index in wheat (Herrmann et al. 2010). CRSI has been documented to be related to salinity in other contexts (Scudiero et al. 2015). Salinity is affected by many factors with drainage and hydraulic conductivity being two important

factors (Eilers et al. 1997). Both factors vary within the different parent material classes in this study, which explains why they were less important for parent material mapping compared to other features.

For the *BSCI-excluded model*, Sentinel 1 backscatter, standard deviation of elevation, and NDVI attributes were most important, with the standard deviation of NDVI as most important NDVI feature followed by the median October NDVI (Table 3). As bare soil composite data were not present, the model found features related to land use such as the standard deviation of NDVI to be the best as separating parent material classes. Sentinel 1 VH backscatter was the most important feature followed by the standard deviation of 9 × 9 median filtered elevation with a 101 × 101 focal window, and then Sentinel 1 VV backscatter. Other important features were ARI, REIP, the standard deviation of 3 × 3 median filtered elevation with a 21 × 21 focal window, median July NDVI, and then the standard deviation of 9 × 9 median filtered elevation with a 21 × 21 focal window, followed by the CRSI. For all three models, the coarsest scale terrain attribute was the most important terrain attribute, as the greatest differences amongst parent material classes were apparent in the larger-scale terrain variability-associated features. Other studies have found local and landscape morphometric features along with hydrological characteristics useful (Heung et al. 2017). With a higher quality DEM, these features would also likely be useful in the Canadian Prairies as well; however, with the 30 m DEM available, the coarser scale properties were more useful in the context of this study. This could be because the scale of terrain variation in the Canadian prairies is such that a 30 m DEM cannot consistently detect all slope positions in the landscape, and the coarser scale terrain analysis has enough of the upper and lower slope positions in the DEM to better characterize overall variability if not specific slope positions at a given point.

Conclusion

Parent material disaggregation has an important role in predictive soil mapping efforts in the Canadian Prairies. The major drivers of soil variation at scales finer than the existing soil maps are parent material variance, slope position, and salinity. Disaggregating parent material maps to create finer resolution maps is an important step for creating more detailed soil maps in the Canadian Prairies to support a variety of end uses. Based on these results, the inclusion of BSCI is an important covariate for parent material disaggregation in the Canadian Prairies and is likely essential for generating useful maps when high-resolution DEMs are not available. The model that did not include BSCI were less accurate overall. Therefore, parent material mapping in the Canadian prairies through disaggregation methods may currently need to be limited to areas where BSCI is available to ensure output maps with acceptable accuracy. Future work to disaggregate soil classes based on slope position and salinity, and to combine those methods with parent material disaggregation is needed to generate detailed soil maps for the Canadian Prairies.

Acknowledgements

We would like to acknowledge the Natural Sciences and Engineering Council of Canada (NSERC) for providing financial support for this project via a post-doctoral fellowship to Dr. Preston Sorenson.

Article information

History dates
Received: 20 October 2021
Accepted: 7 July 2022
Accepted manuscript online: 15 July 2022
Version of record online: 23 November 2022

Notes
This paper is part of a Collection entitled “Advances in Soil Survey & Classification in Canada”.

Data availability
The data used for this project are publicly available from the Canadian Soil Information Service (Agriculture and Agri-Food Canada 2016). All covariate data and analyses can be generated using open source code made available on GitHub with a CC0 1.0 Universal License (Sorenson 2021*a*, 2021*b*, 2021*c*).

Author information

Author ORCIDs
P.T. Sorenson https://orcid.org/0000-0002-2958-1246
J. Kiss https://orcid.org/0000-0001-8195-1518

Author notes
A.K. Bedard-Haughn served as a Guest Editor at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by Chuck Bulmer.

Author contributions
Preston Sorenson: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.
Jeremy Kiss: Writing – review & editing.
Angela Bedard-Haughn: Writing – review & editing.

Competing interests
The authors declare there are no competing interests.

References

Agriculture and Agri-Food Canada. 2016. National Pedon Database. [Online] Available from https://open.canada.ca/data/en/dataset/6457fad6-b6f5-47a3-9bd1-ad14aea4b9e0 [accessed 17 February 2021].

Agriculture and Agri-Food Canada. 2019. Annual crop inventory. [Online] Available from https://open.canada.ca/data/en/dataset/ba2645d5-4458-414d-b196-6303ac06c1c9 [accessed 1 February 2021].

Anderson, D., and Cerkowniak, D. 2010. Soil formation in the Canadian prairie region. Prairie Soils Crop. **3**: 57–64. [Online] Available from www.prairiesoilsandcrops.ca.

Bartholomeus, H., Epema, G., and Schaepman, M. 2007. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. **9**: 194–203. doi:10.1016/j.jag.2006.09.001.

Bartholomeus, H.M.M., Schaepman, M.E.E., Kooistra, L., Stevens, A., Hoogmoed, W.B.B., and Spaargaren, O.S.P.S.P. 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma, **145**: 28–36. doi:10.1016/j.geoderma.2008.01.010.

Becker, W.R., Ló, T.B., Johann, J.A., and Mercante, E. 2021. Statistical features for land use and land cover classification in Google Earth Engine. Remote Sens. Appl. Soc. Environ. **21**: 100459. Elsevier B.V. doi:10.1016/j.rsase.2020.100459.

Behrens, T., Zhu, A.-X., Schmidt, K., and Scholten, T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, **155**: 175–185. Elsevier B.V. doi:10.1016/j.geoderma.2009.07.010.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., and Edwards, T.C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, **239–240**: 68–83. Elsevier B.V. doi:10.1016/j.geoderma.2014.09.019.

Bulmer, C., Schmidt, M.G., Heung, B., Scarpone, C., Zhang, J., Filatow, D., et al. 2016. Improved soil mapping in British Columbia, Canada, with legacy soil data and random forest. *In* Digital soil mapping across paradigms, scales and boundaries. *Edited by* G.L. Zhang, D. Brus, F. Liu, X.D. Song and P. Lagacherie. Springer, Singapore. pp. 291–303. doi:10.1007/978-981-10-0415-5_24.

Burrough, P.A., van Gaans, P.F.M., and Hootsmans, R. 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma, **77**: 115–135. doi:10.1016/S0016-7061(97)00018-9.

Caglar, B., Becek, K., Mekik, C., and Ozendi, M. 2018. On the vertical accuracy of the ALOS world 3D-30 m digital elevation model. Remote Sens. Lett. **9**: 607–615. doi:10.1080/2150704X.2018.1453174.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., and Odgers, N.P. 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma, **274**: 54–67. Elsevier B.V. doi:10.1016/j.geoderma.2016.03.025.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., and Gerlitz, L., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. **8**: 1991–2007. doi:10.5194/gmd-8-1991-2015.

DeLancey, E.R., Kariyeva, J., Bried, J.T., and Hird, J.N. 2019. Large-scale probabilistic identification of boreal peatlands using Google Earth Engine, open-access satellite data, and machine learning. PLoS ONE, **14**: 1–23. doi:10.1371/journal.pone.0218165.

Demattê, J.A.M., Fongaro, C.T., Rizzo, R., and Safanelli, J.L. 2018. Geospatial soil sensing system (GEOS3): a powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. **212**: 161–175. Elsevier. doi:10.1016/j.rse.2018.04.047.

Demattê, J.A.M., Safanelli, J.L., Poppiel, R.R., Rizzo, R., Silvero, N.E.Q., Mendes, W. de S., et al. 2020. Bare earth's surface spectra as a proxy for soil resource monitoring. Sci. Rep. **10**: 1–11. doi:10.1038/s41598-020-61408-1.

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., and Li, X. 2016. 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. **8**: 354. doi:10.3390/rs8040354.

Eilers, R.G., Eilers, W.D., and Fitzgerald, M.M. 1997. A salinity risk index for soils of the Canadian prairies. Hydrogeol. J. **5**: 68–79. doi:10.1007/s100400050118

Gitelson, A.A., Merzlyak, M.N., and Chivkunova, O.B. 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. Photochem. Photobiol. **74**: 38. doi:10.1562/0031-8655(2001)074⟨0038:OPANEO⟩2.0.CO;2.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. **202**: 18–27. doi:10.1016/j.rse.2017.06.031.

He, Y., Wei, Y., DePauw, R., Qian, B., Lemke, R., Singh, A., et al. 2013. Spring wheat yield in the semiarid Canadian prairies: effects of precipitation timing and soil texture over recent 30 years. F. Crop. Res. **149**: 329–337. Elsevier B.V. doi:10.1016/j.fcr.2013.05.013.

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., and Bonfil, D.J. 2010. Assessment of leaf area index by the red-edge inflection point derived from VENμS bands. Hyperspectral 2010 Work. **2010**: 17–19.

Heung, B., Bulmer, C.E., and Schmidt, M.G. 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma, **214–215**: 141–154. doi:10.1016/j.geoderma.2013.09.016.

Heung, B., Hodúl, M., and Schmidt, M.G. 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma, **290**: 51–68. doi:10.1016/j.geoderma.2016.12.001.

Hird, J., DeLancey, E., McDermid, G., and Kariyeva, J. 2017. Google Earth Engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping. Remote Sens. **9**: 1315. doi:10.3390/rs9121315.

Holmes, K.W., Griffin, E.A., and Odgers, N.P. 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. Soil Res. **53**: 865. doi:10.1071/SR14270.

Japan Aerospace Exploration Agency. 2015. ALOS Global Digital Surface Model. [Online]. Available from https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm [accessed 24 September 2021].

Jenny, H. 1941. Factors of soil formation. Soil Sci. **52**: 415. doi:10.1097/00010694-194111000-00009

Kiss, J., and Bedard-Haughn, A. 2021. Predictive mapping of solute-rich wetlands in the Canadian prairie pothole region through high-resolution digital elevation model analyses. Wetlands, **41**: 38. doi:10.1007/s13157-021-01436-3.

Kokaly, R.F., Clark, R.N., Swayze, G.A., Livo, K.E., Hoefen, T.M., Pearson, N.C., et al. 2017. USGS Spectral Library Version 7. Reston, VI. doi: http://dx.doi.org/10.3133/ds1035.

Landi, A., Mermut, A.R., and Anderson, D.W. 2004. Carbon distribution in a hummocky landscape from Saskatchewan, Canada. Soil Sci. Soc. Am. J. **68**: 175–184. doi:10.2136/sssaj2004.1750.

Laurin, G.V., Balling, J., Corona, P., Mattioli, W., Papale, D., Puletti, N., et al. 2018. Above-ground biomass prediction by Sentinel-1 multitemporal data in Central Italy with integration of ALOS2 and Sentinel-2 data. J. Appl. Remote Sens. **12**: 1. doi:10.1117/1.JRS.12.016008.

Malone, B.P., Minasny, B., and McBratney, A.B. 2017. Using R for digital soil mapping. Springer International Publishing, Cham. doi:10.1007/978-3-319-44327-0.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B. 2003. On digital soil mapping. Geoderma, **117**: 3–52. doi:10.1016/S0016-7061(03)00223-4.

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Iversen, B.V., Greve, M.H., and Minasny, B. 2019. Improved disaggregation of conventional soil maps. Geoderma, **341**: 148–160. Elsevier. doi:10.1016/j.geoderma.2019.01.038.

Odgers, N.P., McBratney, A.B., and Minasny, B. 2011. Bottom-up digital soil mapping. II. Soil series classes. Geoderma, **163**: 30–37. Elsevier B.V. doi:10.1016/j.geoderma.2011.03.013.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., and Clifford, D. 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma, **214–215**: 91–100. Elsevier B.V. doi:10.1016/j.geoderma.2013.09.024.

Pennock, D., Bedard-Haughn, A., Kiss, J., and van der Kamp, G. 2014. Application of hydropedology to predictive mapping of wetland soils in the Canadian prairie pothole region. Geoderma, **235–236**: 199–211. Elsevier B.V. doi:10.1016/j.geoderma.2014.07.008.

Plante, A.F., Conant, R.T., Stewart, C.E., Paustian, K., and Six, J. 2006. Impact of soil texture on the distribution of soil organic matter in physical and chemical fractions. Soil Sci. Soc. Am. J. **70**: 287–296. doi:10.2136/sssaj2004.0363.

Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T., and Heiden, U. 2018. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). Remote Sens. Environ. **205**: 1–17. Elsevier. doi:10.1016/j.rse.2017.11.004.

Rossel, R.A.A.V., Behrens, T., Viscarra Rossel, R.A., and Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, **158**: 46–54. Elsevier B.V. doi:10.1016/j.geoderma.2009.12.025.

Safanelli, J.L., Chabrillat, S., Ben-Dor, E., and Demattê, J.A.M. 2020. Multispectral models from bare soil composites for mapping topsoil properties over Europe. Remote Sens. **12**: 1369. doi:10.3390/RS12091369.

Scudiero, E., Skaggs, T.H., and Corwin, D.L. 2015. Regional-scale soil salinity assessment using Landsat ETM + canopy reflectance. Remote Sens. Environ. **169**: 335–343. Elsevier B.V. doi:10.1016/j.rse.2015.08.026.

SKSIS Working Group 2018. Saskatchewan Soil Information System—SKSIS. [Online] Available from sksis.usask.ca [accessed 3 May 2021].

Sorenson, P. 2021a. Google Earth Engine scripts for generating predictive soil mapping environmental covariates. [Online] Available from https://github.com/prestonsorenson/Google_Earth_Engine_PSM/tree/main [accessed 16 August 2021].

Sorenson, P. 2021b. Landsat 5 bare soil composite script. [Online] Available from https://github.com/prestonsorenson/GEE_Bare_Soil_Composite/blob/main/Bare_Soil_Composite_Landsat_5 [accessed 30 April 2021].

Sorenson, P. 2021c. Parent material disaggregation scripts. [Online] Available from https://github.com/prestonsorenson/Parent_Material_Disaggregation/tree/main [accessed 19 August 2021].

Sorenson, P.T., Shirtliffe, S.J., and Bedard-Haughn, A.K. 2021. Predictive soil mapping using historic bare soil composite imagery and legacy soil survey data. Geoderma, **401**: 115316. Elsevier B.V. doi:10.1016/j.geoderma.2021.115316.

Statistics Canada 2015. Percentage of total land prepared for seeding, 1991 and 2006. [Online] Available from https://www150.statcan.gc.ca/n1/pub/16-002-x/2008003/tables/5212582-eng.htm [accessed 1 February 2021].

Statistics Canada 2017. 2016 Census of Agriculture. [Online] Available from https://www.statcan.gc.ca/eng/ca2016 [accessed 27 August 2020].

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C., and Strauss, P. 2018. Sensitivity of Sentinel-1 backscatter to vegetation dynamics: an Austrian case study. Remote Sens. **10**: 1–19. doi:10.3390/rs10091396.

Wright, M.N., and Ziegler, A. 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. **77**: 1–17. doi:10.18637/jss.v077.i01.