Provincial-scale digital soil mapping using a random forest approach for British Columbia

Authors: Heung, Brandon, Bulmer, Chuck E., Schmidt, Margaret G., and Zhang, Jin

Source: Canadian Journal of Soil Science, 102(3) : 597-620

Published By: Canadian Science Publishing

URL: https://doi.org/10.1139/cjss-2021-0090

Canadian Journal of Soil Science

OPEN ACCESS | Article

Provincial-scale digital soil mapping using a random forest approach for British Columbia

Brandon Heung[a], Chuck E. Bulmer[b], Margaret G. Schmidt[c], and Jin Zhang[d]

[a]Department of Plant, Food, and Environmental Sciences, Faculty of Agriculture, Dalhousie University, 21 Cox Road, Truro, NS B2N 5E3, Canada; [b]British Columbia Ministry of Forests, Lands, Natural Resource Operations and Rural Development, Vernon, BC V1B 1S6, Canada; [c]Department of Geography and School of Environmental Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada; [d]Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

Corresponding author: **Brandon Heung** (email: Brandon.Heung@dal.ca)

Abstract

Although British Columbia (BC), Canada, has a rich history of producing conventional soil maps (CSMs) between 1925 and 2000, the province still lacks a detailed soil map with a comprehensive coverage due to the cost and time required to develop such a product. This study builds on previous digital soil mapping (DSM) research in BC and develops provincial-scale maps. Soil taxonomic classes (e.g., great groups and order) and parent material classes were mapped at a 100 m spatial resolution for BC (944 735 km^2). Training points were generated from detailed and semi-detailed soil survey maps. The training points were intersected with 26 topographic indices for mapping parent materials with an additional 9 climatic and vegetation indices for mapping soil classes. The soil–environmental relationships were inferred using the random forest (RF) classifier. The fitted models were used to predict 23 soil great groups, 9 soil orders, and 10 parent material classes. Accuracy assessments were performed using $n = 14\ 570$ validation points for parent material classes and $n = 14\ 316$ validation points for soil classes, acquired from the BC Soil Information System. The accuracy rates for soil great groups, orders, and parent material classes were 55%, 62%, and 69%, respectively, and kappa coefficients were 0.37, 0.41, and 0.59, respectively. This study demonstrated that when RF was trained using CSMs, the accuracy for the resulting DSM was higher than the original CSM. To assess prediction uncertainties, ignorance uncertainty maps were developed using class-probability layers generated by the RF models.

Key words: digital soil mapping, machine learning, soil classification, soil parent material, random forest

Résumé

Bien qu'elle ait produit une foule de cartes pédologiques ordinaires entre les années 1925 et 2000, la Colombie-Britannique, une province canadienne, n'a toujours pas de carte générale détaillée de ses sols, faute de temps et d'argent pour la réaliser. Dans le cadre de cette étude, les auteurs se sont fondés sur des recherches antérieures en cartographie numérique du sol pour tracer des cartes pédologiques à l'échelle de la province. Les classes taxonomiques de sol (groupes principaux et ordres) et les classes de matériau originel ont été reportées sur une carte de la province (944 735 km^2) d'une résolution spatiale de 100 m. Des points d'apprentissage ont ensuite été produits à partir de levés pédologiques précis ou semi-détaillés, et les auteurs les ont entrecoupés avec 26 indices topographiques pour cartographier le matériau originel et avec neuf indices sur le climat et la végétation pour cartographier les classes de sol. Les liens entre le sol et l'environnement ont été déduits par la technique de classification de la forêt aléatoire. Après ajustement, les modèles ont servi à prévoir 23 grands groupes de sol, neuf ordres de sol et dix classes de matériau originel. Les auteurs ont évalué l'exactitude du modèle grâce à des points de validation tirés de la base de données provinciale sur les sols, soit 14 570 points pour les classes de matériaux originel et 14 316 points pour les classes de sol. Le degré d'exactitude pour les grands groupes de sol, les ordres et les classes de matériau originel s'établissait respectivement à 55 %, à 62 % et à 69 %, avec un coefficient kappa correspondant de 0,37, 0,41 et 0,59. Cette étude montre que la formation du modèle de la forêt aléatoire avec des cartes pédologiques ordinaires permet d'obtenir des cartes numériques plus précises que les cartes originales. Les auteurs ont produit des cartes d'incertitude de l'ignorance pour évaluer l'incertitude des prévisions en recourant aux couches de probabilité des classes obtenues avec le modèle de la forêt aléatoire. [Traduit par la Rédaction]

Mots-clés : carte numérique des sols, apprentissage automatique, classification des sols, matériau originel, forêt aléatoire

Introduction

There is an increasing need for the Province of British Columbia (BC), and elsewhere in the world, to provide reliable information on soil patterns to help define and realize the benefits of sustainable resource use and to ensure environmental protection (Carré et al. 2007; Hartemink and McBratney 2008). A close linkage between soil properties and plant productivity underpins the application of precise and accurate soil information for the purposes of forest harvest planning (Brown 1973; Valentine 1986), agricultural production (Bertrand et al. 1991), and grazing management (Teague et al. 2011). The position of soil — at the physical interface between the lithosphere, atmosphere, and hydrosphere — accounts for its control over hydrologic processes, such as rainfall interception, and the partitioning of the received water between infiltration and surface runoff (Wilding and Lin 2006). These landscape processes have a key role in efforts to improve watershed management, protection, and restoration and our ability to respond to natural disasters, such as flooding and drought. Furthermore, the fundamental role of the soil in ecosystem processes provides the context for soil information as a critical component of efforts to respond to natural or anthropogenic disturbances, such as wildfire, insect outbreak, and climate change (Lal 2004).

According to Anderson and Smith (2011) and McKeague and Stobbe (1978), soil survey activities in BC started in the 1920s. An early example of a conventional soil map (CSM) for BC was produced by Kelley and Spilsbury (1939). Soil Survey Report Number 1 provided descriptions of the soils for the Lower Fraser Valley region of the province, and their association with the climatic, topographic, and economic factors that affected agricultural potential. The report contained a detailed description on the trend of agricultural development and settlement in the survey area, and where the most fertile areas were. The identification of areas suitable for farming was the main purpose of CSM in those years. Over the next five decades, more than 100 survey projects were completed in BC. The production of that series of extensive CSMs ended once the major areas with agriculture potential were mapped. One of the last extensive projects was prepared for the Stikine–Iskut area by Fenger and Kowall (1992). Since the late 1990s, some soil maps have been produced for smaller areas, and usually in response to a specific resource management project or information need. Currently, most of the digitized CSM reports are readily available through the Canadian Soil Information System (CanSIS) (Agriculture and Agri-Food Canada 2021).

To leverage the revolutionary advances in the acquisition of environmental data, computational ability, and geographic information systems (GIS), the application of digital soil mapping (DSM) techniques has steadily progressed over the last decade in BC. Early projects included the development of predictive ecological maps for a portion (3 million ha) of the Cariboo Forest Region of central BC, using fuzzy classification systems (MacMillan et al. 2007) — an approach that was later extended to the remainder of the region (MacMillan et al. 2010). To further test the applications of fuzzy inference approaches and extend their applicability toward the disaggregation and refinement of legacy CSMs, Smith et al. (2012) proposed the use of a weights-of-evidence technique to inform the process of defining the fuzzy rule sets used in the ArcSIE software (Shi et al. 2009; Shi 2010). In Smith et al. (2012), the soil classes were mapped for the Trout Creek region of the Okanagan Basin and in Smith et al. (2016) DSMs were produced for soil classes and soil attributes for the remainder of the Okanagan Basin in accordance with GlobalSoilMap.net specifications.

In addition to the use of fuzzy classification techniques, machine-learning techniques have been tested for the Lower Fraser Valley (Heung et al. 2014, 2016) and the Okanagan–Kamloops regions (Heung et al. 2017) of BC. In Heung et al. (2014), a framework for data mining the soil–environmental relationships from detailed legacy CSMs was evaluated for the mapping of soil parent material using the random forest (RF) classifier. A similar approach was used by Bulmer et al. (2016) to produce a map of parent material and a limited set of soil classes by mosaicking a series of regional maps. Because of the successful application of the RF classifier, it was subsequently used to produce exposed bedrock maps for the Tulameen–Princeton region of southcentral BC (Scarpone et al. 2017). Furthermore, the RF algorithm was also compared against a generalized linear model and regression kriging approaches for the purposes of mapping soil thickness for that same region (Scarpone et al. 2016).

To further explore the application of machine-learning techniques for producing DSMs in BC, Heung et al. (2016) presented an overview and comparison of a variety of machine-learning techniques for mapping soil classes for the Lower Fraser Valley. In that study, 10 machine-learning techniques were compared, where it was observed that support vector machine with radial basis function and *k*-nearest neighbours produced results with the highest accuracy; however, the classification and regression trees (CART) with bagging, logistic model trees, and RF classifiers performed comparably well. Although the support vector machine produced the best model, Heung et al. (2016) indicated that the learner was computationally demanding and the process of optimizing the hyperparameter values was challenging. With respect to the *k*-nearest neighbours learner, its accuracy was also high in Heung et al. (2016); however, the spatial patterns that were generated from those predictions were inconsistent with the pedological knowledge of the authors.

Further exploration was done for the Okanagan–Kamloops region of BC, where Heung et al. (2017) presented a comparison of DSMs derived from legacy soil pits and CSM polygons for mapping soil classes. There, it was observed that the DSMs produced using CSM polygons as training data were more accurate in comparison to maps produced by using soil pedon data. Furthermore, additional comparisons between single-model learners and ensemble-model learners were performed, where it was observed that ensemble-model learners consistently produced more accurate results with the RF classifier performing consistently well — regardless of the type

of training data used. The effectiveness of the RF classification algorithm was consistent with other model comparison studies within the DSM literature (e.g., Brungard et al. 2015; Taghizadeh-Mehrjardi et al. 2015). Furthermore, RF has the added benefit in being computationally efficient when optimizing its main hyperparameter, m_{try}, whereby the number of potential hyperparameter values is constrained to the number of predictors (Heung et al. 2016). Lastly, the model has the added ability of generating uncertainty maps (Heung et al. 2017).

Despite the progress made in DSM research in BC, many of the developments have been limited to the production of regional maps and as a result, the availability of a high-resolution digital soil data set for the entirety of the province remains limited. Currently, the only uniform, province-wide, soil map is the Soil Landscapes of Canada (SLC) Version 2.2 data set, which was produced at a 1:1 000 000 scale and provided by CanSIS. The SLC data set was created by digitizing a combination of provincial- and regional-scale CSMs and is available in the polygon format (Schut et al. 2011). The portion of the SLC that covers BC consists of 2651 multicomponent polygons with an average polygon size of 380 km^2 (Geng et al. 2010); as such, its usefulness remains limited when spatially explicit and detailed soil information is needed. Hence, the objectives of this study were as follows: (1) to demonstrate and extend the use of the RF classifier toward the development of provincial DSMs of soil parent material classes and soil taxonomic classes (great groups and orders) at a 100 m spatial resolution and (2) to validate the predictions using the BC Soil Information System (BCSIS) data set — a georeferenced repository of soil information.

Materials and methods

The methods used for this study were adopted from previous regional-scale DSM research that was carried out for the mapping of soil taxonomic class (great groups and orders) and parent material for the Lower Fraser Valley (Heung et al. 2014, 2016) and Okanagan–Kamloops regions (Heung et al. 2017) of BC.

Soil-forming environment of British Columbia

British Columbia is the western-most province of Canada, located between 48–60°N and 114–139°W and with an areal extent of 944 735 km^2 (9.5% of Canada; Fig. 1). The peak elevation of the province is located on Fairweather Mountain at 4671 m (above mean sea level). British Columbia lies within the Western Cordillera of North America, and its physiography was described in Church and Ryder (2010) as a broad, complex system of mountains and plateaus. The complex physiography and geology reflect BC's location on the western edge of North America, where for more than 350 million years, numerous geologic terranes have collided with the ancestral continent through the process of plate tectonics (Monger 1997). A series of northwest- to southeast-trending mountain ranges have resulted from the tectonic activity, while intervening periods of erosion and volcanism led to the presence of gently sloping plateaus that dominate large portions of central BC. The geology is highly diverse at local scales, while the broad patterns in rock type reflect the character of the accreted landmasses, as well as the tectonic and geomorphic processes of orogeny, volcanism, metamorphism, erosion, and deposition.

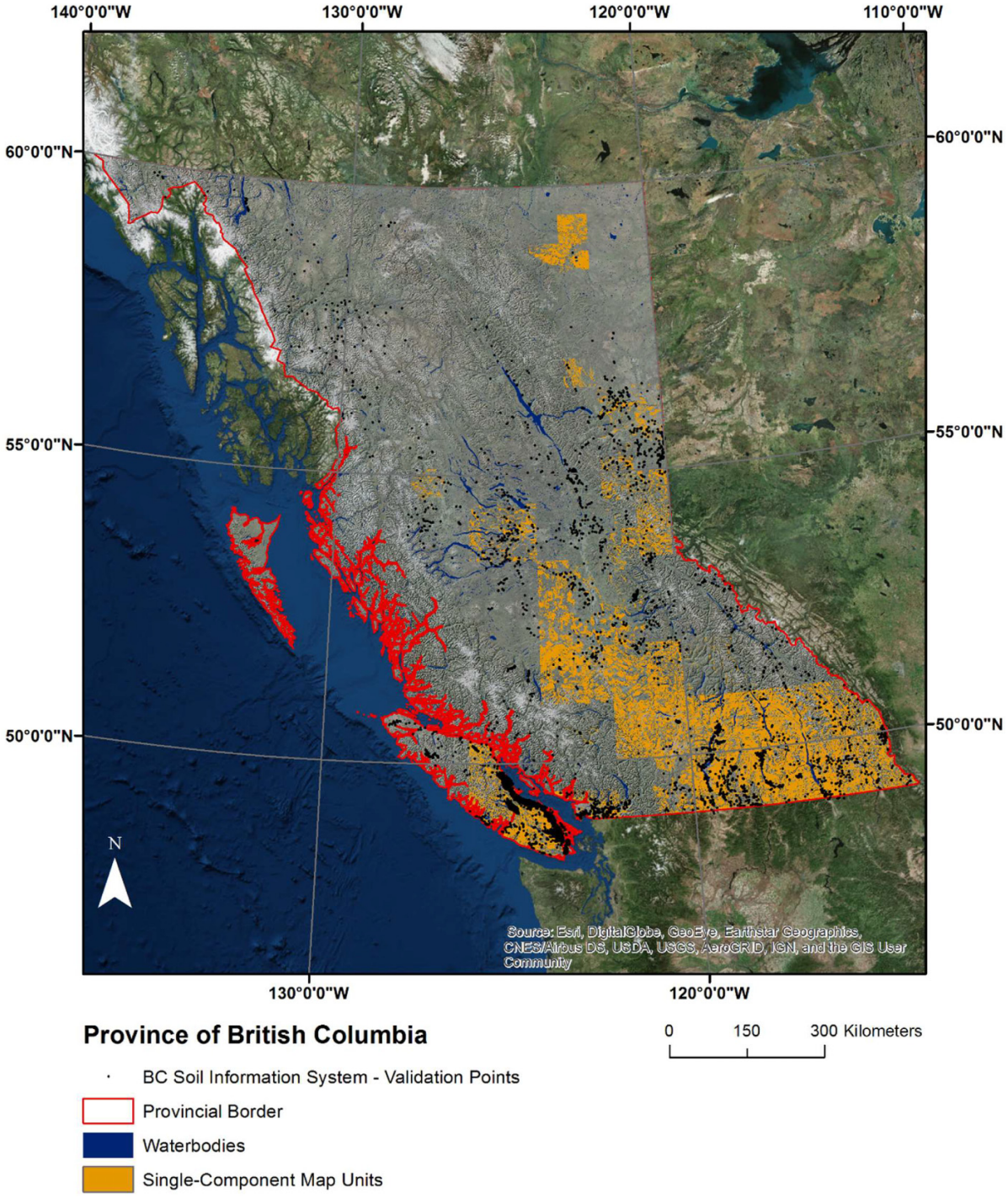
Holland (1976) outlined three broad mountain systems that occupy the majority of BC, in addition to a small portion (10%) of the provincial landmass occupied by the Interior Plains of North America. The western system of mountains includes several ranges where intrusive igneous rocks are the dominant bedrock type. The Interior Plateau has a wide diversity of rock types but includes a large area of flat lying and folded lava known as the plateau basalts. To the east of the Interior Plateau, the Eastern System consists primarily of sediments that were deposited into the ocean off the coast of the ancestral continent and were subsequently uplifted as a result of the tectonic activity. These mountain ranges include the Rocky Mountains, where the continental divide marks the provincial boundary in the southeast. The Eastern System adjoins the Alberta Plateau in the northeastern part of the province, is dominated by sedimentary rocks, and is generally flat in topography.

Church and Ryder (2010) provide a description of landforming processes in relation to the spatial scale of the resulting landforms. The large landscape features (observable at scales of >100 km) and bedrock characteristics result from the tectonic processes of orogenesis and volcanism associated with the movement of continental plates, punctuated by long intervening periods of erosion. For at least the past 2 000 000 years, glaciers advanced and receded throughout BC with the last advance ending approximately 10 000 years B.P. The influence of ice in shaping the broad landscape is observable at scales up to 1000 km, while the contemporary geomorphic processes of erosion, landslides, flooding, and deposition continue to shape the land into features visible on maps at finer scale. Soil-forming processes within pedons can be considered to operate at localized scales and influence the properties of the surface to a depth of ~1 m within the unconsolidated surficial (parent) material.

Because of the highly variable topography of the Coast and Rocky Mountain Ranges and the proximity to the Pacific Ocean, BC experiences a combination of maritime and continental climate patterns. The wettest parts of BC are located along the Pacific Coast and especially on the windward slopes of Vancouver Island, Haida Gwaii (formerly the Queen Charlotte Islands) and the Coast Mountains as a result of orographic precipitation. The driest parts of the province are in southcentral BC, behind the rain shadow and leeward of the Coast Mountains. In the Great Plains region of northeastern BC, the continental climate results in the coldest temperatures in BC. A system of biogeoclimatic ecosystem classification (BEC) based on climate and vegetation type is used in BC to describe the ecological character of 14 biogeoclimatic zones (Pojar et al. 1991).

The Coastal Western Hemlock (CWH) biogeoclimatic zone has abundant rainfall, with a mean annual precipitation of 2228 mm, which ranges from 1000 to 4400 mm across the province. Here, the vegetation is comprised of a mixture of western hemlock (*Tsuga heterophylla*), Douglas-fir (*Pseudotsuga menziesii*), and western redcedar (*Thuja plicata*) tree species.

Fig. 1. Province of BC with the spatial distribution of single-component map units (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018) and validation points acquired from the BC Soil Information System (Sondheim and Suttie 1983; BC Albers projection). [Colour online]



The soils of the CWH zone are typically Humo-Ferric Podzols, which transition to Ferro-Humic Podzols with increasing precipitation and decreasing temperatures (Pojar et al. 1991). In southcentral BC, the Bunchgrass (BG) zone is the warmest and driest biogeoclimatic zone in BC, with a mean annual temperature of 10.0 °C and 242 mm of mean annual precipitation. The soils of the BG zone are typically Brown Chernozems, which transition to Dark Brown and Black Chernozems with decreasing temperatures (Nicholson et al. 1991). The Boreal White and Black Spruce (BWBS) zone in northeast BC has a continental climate and experiences a mean annual temperature of −1.4 °C; furthermore, this zone is also dry with a mean annual precipitation of 327 mm. The vegetation in the BWBS is comprised of predominantly black spruce (*Picea mariana*), which transitions to a mixture of white spruce (*Picea glauca*) and trembling aspen (*Populus tremuloides*) with increas-

ing elevation. For the well-drained areas of the BWBS zone, the soils are mostly Gray Luvisols, while in the poorly drained lowlands, Organic Cryosols, Luvic Gleysols, and Organic soils are common. Where soils are developed from lacustrine clays over marine shale (e.g., Dawson Creek and Fort St. John), Solonetzic soils may be found (DeLong et al. 1991).

Environmental variables

A suite of 35 environmental variables were derived from a combination of land cover classification, climate, digital elevation, and ecosystem classification data sets — all of which were resampled to a 100 m spatial resolution (Table 1). For mapping soil parent material, only topographic indices were used because in Heung et al. (2014) and Bulmer et al. (2016) it was observed that topographic indices were well correlated with the distribution of soil parent material. Given the glacial history of BC, many of the soil parent materials were transported across the landscape, which then resulted in the distinct topographic features that were produced due to the deposition of sediments. For mapping soil classes, vegetation and climatic indices were included in addition to the topographic indices.

Topographic indices

Topographic indices were mostly derived from BC's Terrain Resource Information Management (TRIM) digital elevation model (DEM) (B.C. Ministry of Sustainable Resource Management 2002). The DEM was originally developed from a triangulated irregular network, derived from TRIM mass-points and break-lines, and aggregated to a 100 m spatial resolution. The 100 m DEM was acquired from HectaresBC.org (HectaresBC 2012), which provides a repository of freely available digital data layers for BC. To minimize the effects of spatially noncorrelated noise on the calculation of topographic indices, the original DEM was filtered using three consecutive mean filters of 3 × 3, 3 × 3, and 5 × 5 pixels (MacMillan et al. 2007).

All DEM-derived topographic indices were calculated in the System for Automated Geoscientific Analysis (SAGA) (SAGA Development Team 2011). The topographic indices were selected based on their ability to represent local-scale morphometry (e.g., slope, aspect, and curvature), landscape-scale morphometry (e.g., slope position and multiresolution valley bottom flatness index), hydrologic characteristics (e.g., slope-length factor and topographic wetness index), and landscape exposure (e.g., skyview factor and visible sky factor).

In addition, two distance-based metrics were included as predictors: distance-to-nearest-stream and distance-to-nearest-lake. The distance-to-nearest-stream layer was calculated from polyline mapping from HectaresBC.org, while the distance-to-nearest-lake layer was calculated from the Freshwater Atlas for BC data set (Integrated Land Management Bureau 2010). The distance-to-nearest-stream was used because it was previously found to be an important predictor when mapping soils developed from fluvial sediments (Heung et al. 2014). The distance-to-nearest-lake was included for this study due to the potential importance of the predictor for mapping soils developed from lacustrine sediments as well as the mapping of hydromorphic soils.

Vegetation indices

To represent the vegetation patterns of the province, the GeoBase Land Cover Circa 2000 (LCC2000) product was included. The LCC2000 data layer was derived from the classification of Landsat 5 and Landsat 7 orthoimagery, produced at a 30 m spatial resolution (Olthof et al. 2009), and aggregated to a 100 m spatial resolution using a nearest neighbour approach. The use of interpreted remote sensing products as a covariate in DSM has been done in multiple studies, such as Collard et al. (2014), which used the Corine Land Cover 2006 data sets to map the soils of northern France; Cheney et al. (2016), which used the 2006 National Land Cover Database to map soil properties of continental United States; and Hengl et al. (2017), which used the GlobCover30 data set to produce a global soil map product.

In addition, digitized BEC zone and subzone maps were included to provide information on vegetation patterns and ecological characteristics. Each of the 14 BEC zones is associated with a distinct climax plant species, where subzones are further delineated based on broad-scale moisture and temperature characteristics (Meidinger and Pojar 1991). The BEC zone and subzone maps were originally in the polygon format but were rasterized to a 100 m spatial resolution using the ArcGIS 10.7.1 software.

Climate indices

Climate data were acquired from Climate BC (Wang et al. 2012) and accessed through HectaresBC.org. Mean annual temperature, mean annual precipitation, degree days <0 °C, climatic moisture deficit, reference evaporation, and frost-free days were included as climatic indices. In addition, the provincial BEC maps describing patterns of vegetation and ecology were closely associated with observed patterns of climate variability in BC; hence, the BEC data were also a climate-related predictor.

Development of training data

Based on previous research, it was concluded that the most appropriate modelling approach would be to develop a training data set using CSM data via the sampling of single-component map units using an area-weighted approach. In Heung et al. (2014), three methods of sampling the training areas were tested for the prediction of soil parent materials for the Lower Fraser Valley: (1) a by-polygon approach, where a set number of sample points were obtained from each polygon; (2) an equal-class approach, where a set number of sample points were obtained from each class; and (3) an area-weighted approach, where the number of sample points per class was proportional to the areal extent of each class. Heung et al. (2016) later sought to examine the effects of class imbalance on classification and in addition to including the three sampling methods from Heung et al. (2014), an area-weighted approach with random oversampling (ROS) was in-

Table 1. List of 35 topographic, climatic, and vegetation predictors at a 100 m spatial resolution.

Representation	Environmental covariate	Code
Local-scale morphometry	Elevation	elev
	Plan curvature	plan
	Profile curvature	prof
	Slope	slope
.	Tangential curvature	tancur
	Terrain ruggedness index	tri
	Transformed aspect	aspect
	Total curvature	curve
Landscape-scale morphometry	Multiresolution ridge top flatness index	mrrtf
	Multiresolution valley bottom flatness index	mrvbf
	Mid-slope position	midslope
	Normalized height	nheight
	Slope height	sheight
	Valley depth	vdepth
Hydrologic characteristics	Convergence index	conv
	Distance-to-nearest-lake	dist2lake
	Distance-to-nearest-stream	dist2river
	Modified relative hydrologic slope position	mrhsp
	Relative hydrologic slope position	rhsp
	SAGA wetness index	swi
	Slope-length factor	ls
	Stream power index	spi
	Topographic wetness index	twi
Landscape exposure	Skyview factor	skyview
	Visible sky factor	vissky
	Terrain view factor	terview
Climatic indices[a]	Degree days < 0 °C	dd0
	Frost-free days	ffd
	Hargreaves climatic moisture deficit	moistdef
	Hargreaves reference evaporation	evap
	Mean annual precipitation	map
	Mean annual temperature	mat
Vegetation indices[a]	Biogeoclimatic ecosystem classification — subzone	bec_subzone
	Biogeoclimatic ecosystem classification — zone	bec_zone
	Canadian Land Cover Classification Circa 2000	circa

[a]Climatic and vegetation indices were not included as variables for soil parent material maps.

cluded. Additionally, Bulmer et al. (2016) compared the area-weighted method to a sampling approach that integrated expert knowledge in the sampling procedure.

Among the sampling methods, it was shown that the area-weighted approach performed consistently better than the other sampling approaches except for the area-weighted approach with ROS. Based on Heung et al. (2016), ROS resulted in small improvements in terms of prediction accuracy; however, the procedure was computationally demanding and it was therefore concluded that its implementation was not feasible for large data sets. Furthermore, the choice of using CSMs as training data was deemed more appropriate because in Heung et al. (2017), higher accuracies were achieved when using CSMs as training data in comparison to pedon data for the Okanagan–Kamloops region.

Digitized CSMs, in the polygon format, were obtained from the BC Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The digitized data consisted of 113 CSMs, which were originally mapped at scales between 1:20 000 and 1:125 000. In BC, the large-scale CSMs were typically developed for agriculturally intensive regions of the province, including the Lower Fraser Valley and the Okanagan Valley, whereas the small-scale CSMs were typically developed for areas with high relief, forest vegetation, and lower populations.

The digitized CSMs consisted of 1035 named soil series and 10 soil parent material classes. To simplify the CSM legend and provide a sufficient number of validation points for validating the accuracy of each individual class, the soil series were aggregated to the soil great group and soil order lev-

els of the Canadian System of Soil Classification (Soil Classification Working Group et al. 1998). The simplification of the CSM legend resulted in 23 soil great groups and 9 soil orders. To reduce the classification uncertainty of the training data, only single-component mapping units of soil great group and parent material were retained as training areas. Within BC, the coverage by single-component map units occupied 12.3% (116 293 km^2) and 11.1% (104 888 km^2) for soil great groups and soil parent material of the province's extent, respectively. It should be noted that when a map unit was described as a "single-component" unit, the CSMs were identifying the dominant soil type or parent materials — other soils may have been present within the polygon but were not identified due to the map scale.

To reduce the size of the training data set and to overcome the computational limitations when training a machine learner, 500 000 points were generated within the single-component polygons using the area-weighted approach described in Heung et al. (2014). Here, the number of points to be generated for each class was weighted by the areal extent of the single-component map units for each class. Table 2 shows the percentage of the 500 000 training points that were generated for each class. The values for each environmental variable were extracted for each training point and the resulting dataframe was then used to train the machine learner. The dominant parent material classes included till and colluvium, which consisted of 47% and 29% of the parent material training data, respectively, while the dominant soil great groups included Humo-Ferric Podzols and Gray Luvisols, which consisted of 33% and 30% of the soil great group training data, respectively (Table 2).

Random forest classifier

Based on previous model comparison studies in BC (Heung et al. 2016, 2017), the RF algorithm was used for this study. RF is based on an ensemble of classification trees that are grown from a randomized bootstrap sample of the training data set (Breiman 2001). The bagging procedure in the RF algorithm is designed to mitigate the impact of high model variance, which is common for tree-based learners (Breiman 2001). The resulting prediction made by the RF ensemble is based on a majority-vote combination function using the individual trees. Furthermore, the algorithm can evaluate variable importance by measuring the mean decrease in Gini (MDG) impurity when an individual variable is removed from the prediction process. If the removal of a particular predictor variable leads to a large decrease in the Gini index, it is assigned a greater importance than predictors whose removal results in a smaller decrease in Gini.

For the purposes of this study, the implementation of the RF model and its parameterization were performed using a combination of the *caret* (Kuhn 2008) and *randomForest* (Liaw and Wiener 2002) packages within the R statistical software (R Development Core Team 2012). The parameterization of the RF models was performed through 20 replicates of fivefold cross-validation as described in Heung et al. (2014).

Assessment of predictions

The accuracy of the predicted soil parent material and soil great group maps was assessed using soil pedon data acquired from the BCSIS database (Sondheim and Suttie 1983). The BCSIS data set is a repository of georeferenced soil observations acquired by multiple provincial and federal agencies for various projects (e.g., terrestrial ecosystem mapping, and habitat, soil, and bioresource monitoring) and compiled by the BC Ministry of Environment. Because most of the data points in the BCSIS were contributed prior to the widespread availability of global positioning systems, the positional accuracy of individual points was variable; furthermore, the positional uncertainty for each sample location was unknown. Due to the potential spatial inaccuracy associated with the use of legacy soil observations, the BCSIS data were cleaned to ensure that only observations with the highest spatial accuracy were used for validation purposes. The data cleaning procedure consisted of four steps: (1) the removal of duplicated soil observations; (2) the removal of observations that were not located within the terrestrial landmass of BC; (3) the removal of observations, where its spatial position was located outside of the project boundary for which it was collected; and (4) the removal of observations, where its field-measured elevation was not within 30 m of the elevation extracted from the provincial DEM.

To further account for spatial inaccuracies of the BCSIS data, as well as to account for situations where a validation point was located along a pixel boundary, predictions were considered to be valid if the validation point matched the prediction within a radius of 1 pixel (i.e., 100 m radius). The BCSIS validation data set used for this study consisted of 14 316 points for soil great groups and orders, and 14 570 points for soil parent materials. To assess soil classes at the "order" level of taxonomy, predictions made at the great group level were aggregated. The by-class distribution of the validation data points for soil taxonomic classes and soil parent material classes is shown in Table 2. To compare the accuracy of the CSMs to the DSMs, the validation points were used to assess the single-component map units, where the validation points that coincided with those map units were then used to assess the accuracy of the DSM. The number of validation points that coincided with the single-component map units were 9336 points for soil great groups and orders, and 9856 points for soil parent materials.

The overall agreement between the validation points and the predictions, C, is calculated as

$$C = \sum_{j=1}^{J} p_{jj} \quad (1)$$

where p is the proportion of correctly classified pixels for the jth class. The overall agreement was used as the main accuracy metric for this study. The overall agreement values were further decomposed into additional accuracy metrics, which included the quantity disagreement (Q) and allocation disagreement (A), using the following relationship:

$$C = 1 - Q - A \quad (2)$$

Table 2. Soil orders, great groups, and parent material classes for BC with corresponding percentages of training and validation data.

Soil taxonomic unit	Order	Order code	Great group	Great group code	Training data (%)	Validation data (%)
	Brunisol	B	Dystric Brunisol	DYB	12.80	32.93
			Eutric Brunisol	EB	11.36	1.91
			Melanic Brunisol	MB	0.12	0.38
			Sombric Brunisol	SB	0.09	5.53
	Chernozem	Ch	Black Chernozem	BLC	0.62	0.87
			Brown Chernozem	BC	0.68	0.17
			Dark Brown Chernozem	DBC	1.00	0.64
			Dark Gray Chernozem	DGC	0.18	0.13
	Cryosol	Cr	Organic Cryosol	OC	0.80	0.00
	Gleysol	G	Gleysol	G	0.24	6.45
			Humic Gleysol	HG	0.43	9.83
			Luvic Gleysol	LG	0.09	1.38
	Luvisol	L	Gray Brown Luvisol	GBL	0.00	0.03
			Gray Luvisol	GL	29.74	1.55
	Organic	O	Fibrisol	F	1.03	0.13
			Folisol	FO	0.00	0.00
			Humisol	H	0.13	2.95
			Mesisol	M	1.36	1.28
	Podzol	P	Ferro-Humic Podzol	FHP	3.42	0.89
			Humo-Ferric Podzol	HFP	33.33	29.52
	Regosol	R	Humic Regosol	HR	0.04	1.08
			Regosol	R	2.46	2.35
	Solonetz	S	Solod	SO	0.08	0.00
Soil parent material	Class	Class code				
	Colluvium	C			28.97	3.92
	Eolian	E			0.22	0.05
	Fluvial	F			6.37	17.98
	Glacio-fluvial	FG			9.06	8.35
	Glacio-lacustrine	LG			3.83	0.47
	Glacio-marine	WG			0.01	0.37
	Lacustrine	L			1.15	1.00
	Marine	W			1.00	40.99
	Morainal/Till	M			45.78	20.95
	Organic	O			3.60	5.92

Both Q and A were calculated from the error matrix used in the validation procedure (Pontius and Millones 2011; Warrens 2015). In eq. 2, Q, calculated as

$$(3) \quad Q = \frac{1}{2}\sum_{i=1}^{j} |p_{i+} - p_{+i}|$$

represents the amount of disagreement in the proportion of each soil class between the prediction and validation data sets. In eq. 2, p_{i+} and p_{+i} represent the row and column totals of the error matrix, respectively, expressed as proportions of the population, for the ith class for j number of soil classes. Values of Q range from 0 to 1, where values near 0 represent high agreement in the proportions of coverage for each class while values near 1 represent high disagreement. In eq. 2, A, calculated as

$$(4) \quad A = \left[\sum_{i=1}^{j} min\left(p_{i+}, p_{+i}\right)\right] - C$$

represents the amount of disagreement in the spatial allocation of classes between the prediction and validation data sets. Values of A range from 0 to 1, where values near 0 represent high agreement in the spatial allocation for each class while values near 1 represent high disagreement.

Lastly, Cohen's kappa coefficient, κ, was also calculated to account for the by-chance agreement between the observed

and predicted classes. The coefficient is formulated as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (5)$$

where $P(A)$ represents the overall agreement fraction and $P(E)$ represents the expected agreement fraction between the observed and predicted soil classes. Given that the by-chance agreement is accounted for in κ, the coefficient is expected to be consistently lower than C. Values of κ range from 0 to 1, where values near 0 represent low agreement while values near 1 represent high agreement between the observed and predicted classes.

Prediction uncertainty

One advantage of using the RF algorithm, and ensemble-modeling techniques in general, is their ability to produce maps of model uncertainty based on the constituent models of the ensemble. For each pixel, the RF algorithm can estimate the probability of occurrence (i.e., proportion of votes) for each class based on the ensemble of decision trees. In Heung et al. (2017), the soil class probability surfaces facilitated a visual assessment of the model results and also identified regions of the landscape where the transition of one soil type to another might occur.

Using the set of class-probability surfaces, ignorance uncertainty surfaces may be produced, which measure how evenly distributed the class-probability surfaces are across all classes. As the class probabilities become more evenly dispersed, the ignorance uncertainty increases; conversely, as the constituent models of the ensemble converge their predictions toward a particular class, the ignorance uncertainty decreases (Leung et al. 1993; Zhu 1997). The ignorance uncertainty is estimated using an entropy measure, H, and was calculated as follows (Zhu 1997):

$$H(x) = -\frac{1}{\ln n}\sum_{k=1}^{n} P_k(x)\ln P_k(x)\ln P_k(x) \quad (6)$$

where P_k is the proportion of instances where pixel x is classified as soil class k and n is the number of classes. Values of H range from 0 to 1, where values near 0 represent low ignorance uncertainty while values near 1 represent high uncertainty in classification. The use of the entropy measure for providing an overall estimation of uncertainty has previously been used for fuzzy inference classification approaches (e.g., Leung et al. 1993; Goodchild et al. 1994; Zhu 1997) and later extended toward ensemble-modelling approaches (e.g., Kempen et al. 2009; Heung et al. 2017; Blackford et al. 2021).

Results and discussion

Assessment of predictions

Provincial-scale digital soil maps of soil parent materials and soil great groups were produced at a 100 m spatial resolution and are shown in Figs. 2 and 3, respectively. Using the BCSIS validation data set, the accuracies for the provincial soil maps were 55% ($\kappa = 0.37$) and 62% ($\kappa = 0.41$) for the soil great group and soil order classes, respectively, and 69% ($\kappa = 0.59$) for the soil parent material class (Table 3). Due to the highly imbalanced class distribution within the training and validation data, the difference between the overall accuracies and the kappa coefficients were quite substantial given the ability of the kappa coefficient to capture class imbalances more effectively.

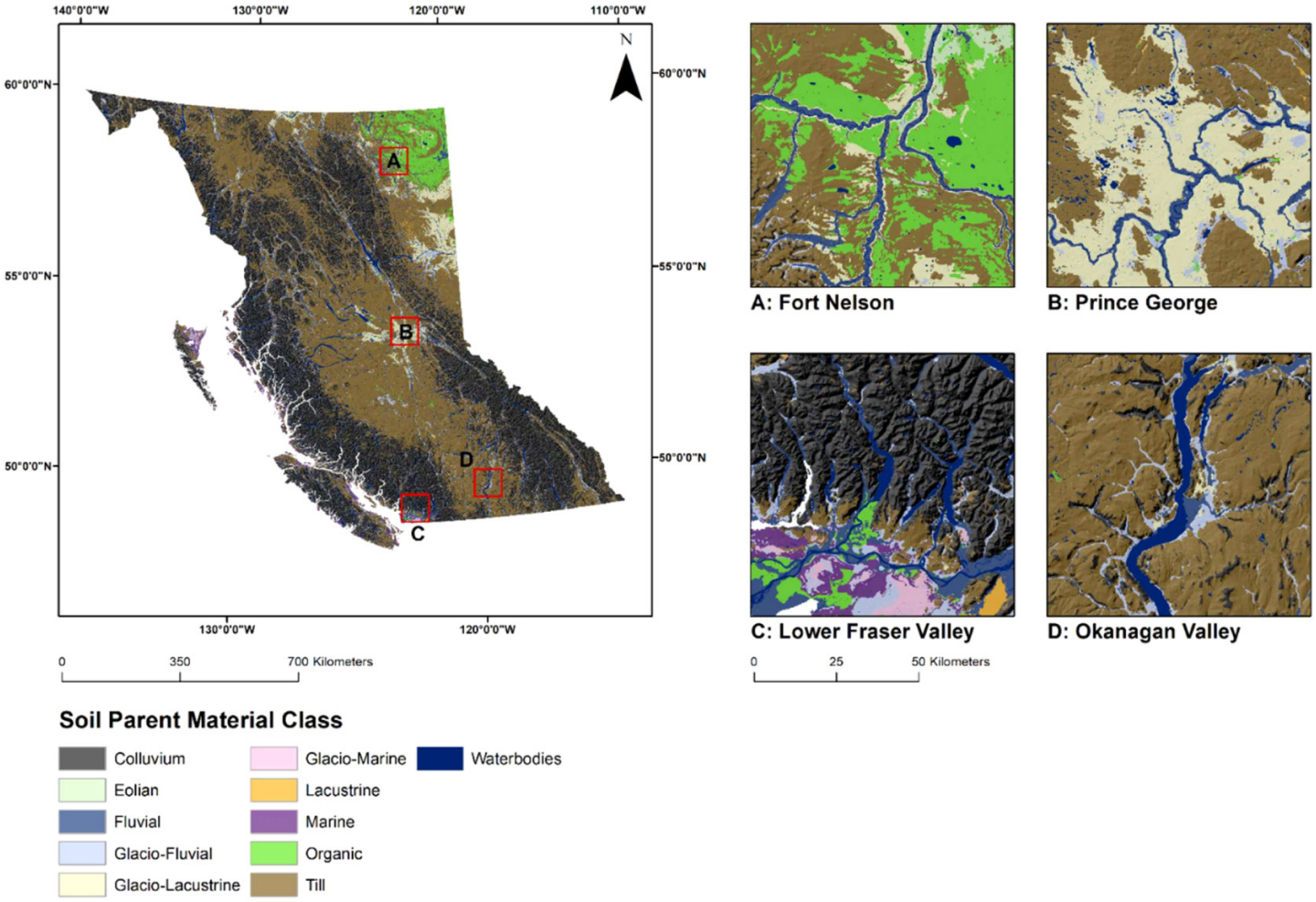
When assessing the accuracy of the single-component map units from the CSM and comparing the accuracy to the predicted pixels that coincided with those map units, it was observed that the DSM approach was able to produce a refined soil map with higher accuracy rates. The DSM approach resulted in increases in accuracy rates by 2% and 6% for soil orders and great groups, respectively, and an increase by 10% for soil parent material classes (Table 3). These results show how the use of CSMs as training data, under a DSM framework, may also serve to refine the original CSM. Previous studies that have shown this effect have included Collard et al. (2014), where a 1:250 000 CSM was improved using various machine-learning approaches; Yang et al. (2011), where a 1:20 000 CSM was improved using a fuzzy inference system and expert knowledge approach; and Heung et al. (2017), where multiple CSMs were improved using a variety of machine-learning approaches. Given the limited number of examples that demonstrate the capacity of DSM approaches in the refinement of CSMs, future DSM research should consider additional comparative analyses.

Soil parent materials

The allocation disagreement accounted for a larger proportion of the overall disagreement in comparison to the quantity disagreement (Table 3). The high accuracy observed for the soil parent material map was most likely due to the abundance and clustering of validation points that were identified as marine sediments. In terms of the training data, the marine class was a minority class and represented 1% of the training data set; yet, it accounted for 41% of the validation data set and an accuracy rate of 79% (Table 2). The clustering occurred primarily along southern Vancouver Island, where extensive soil data were acquired and where marine sediments are common. Glacial till, the majority class, provincially, was well predicted with an accuracy rate of 76%; however, an inspection of the parent material confusion matrix (Table 4) also indicates that the till class was overpredicted at the cost of other classes such as lacustrine, organic, and marine.

The lacustrine and eolian classes were poorly predicted, at least partly, due to a combination of limited training and validation points (Table 2). Within BC, the likelihood of observing a pure eolian parent material is very low because it typically occurs as a 0.1–1.0 m thick veneer over other parent material classes such as glacio-fluvial, glacio-lacustrine, glacio-marine, and till materials. As a result, the eolian materials often have little impact on the overall soil profile characteristics compared to the underlying materials. When compared to pre-

Fig. 2. Soil parent material map using a random forest classifier at a 100 m spatial resolution. The map is shown with an underlying hill-shade with insets for the Fort Nelson (A), Prince George (B), Lower Fraser Valley (C), and Okanagan Valley (D) regions (BC Albers projection). [Colour online]



vious regional-scale mapping efforts in the Lower Fraser Valley region of the province, the accuracy rate of the glacio-marine class (20%) was considerably lower for this study in comparison to the 80% accuracy that was reported in Heung et al. (2014). Overall, the distribution of glacio-marine sediments is closely associated with marine sediments, where glacio-marine sediments are located at higher elevations due to isostatic rebound (Luttmerding 1981); however, given the size and geographic extent of the provincial-scale training data set, the subtle differences in elevation between the two classes may have been masked. Within the CSMs, the occurrence of glacio-marine materials was restricted to the Lower Fraser Valley, while the CSMs for the adjacent region of southern Vancouver Island identified a large distribution of marine materials, which likely would have contributed to the masking of glacio-marine presence in the training data. However, it should also be noted that the detailed soil surveys for Vancouver Island did not differentiate between glacio-marine and marine parent materials, potentially contributing to additional model uncertainty.

Soil taxonomic class

The soil great group and order maps had higher quantity disagreement rates at 27% and 22%, respectively, in comparison to the allocation disagreement (Table 3). The reduction in the quantity disagreement by 5% between the taxonomic levels was likely due to the simplification of the legend from soil great groups ($C = 55\%$) to orders, where the misclassification of validation points between soil great groups within the same order was eliminated. The accuracy of the soil great group map was largely influenced by the effective prediction of Dystric Brunisols, which accounted for 33% of the validation data set (Table 2) and had an accuracy rate of 90% (Table 5). Despite the Dystric Brunisols not being the clear majority class in comparison to Humo-Ferric Podzols that accounted for 33% of the training data, an inspection of the confusion matrix indicates that the Humo-Ferric Podzols were preferentially misclassified as Dystric Brunisols (Table 5). The misclassification between these two great groups was not unexpected because Dystric Brunisols are closely associated with

Fig. 3. Soil great group map using a random forest classifier at a 100 m spatial resolution. The map is shown with an underlying hill-shade with insets for the Fort Nelson (A), Prince George (B), Lower Fraser Valley (C), and Okanagan Valley (D) regions (BC Albers projection). [Colour online]

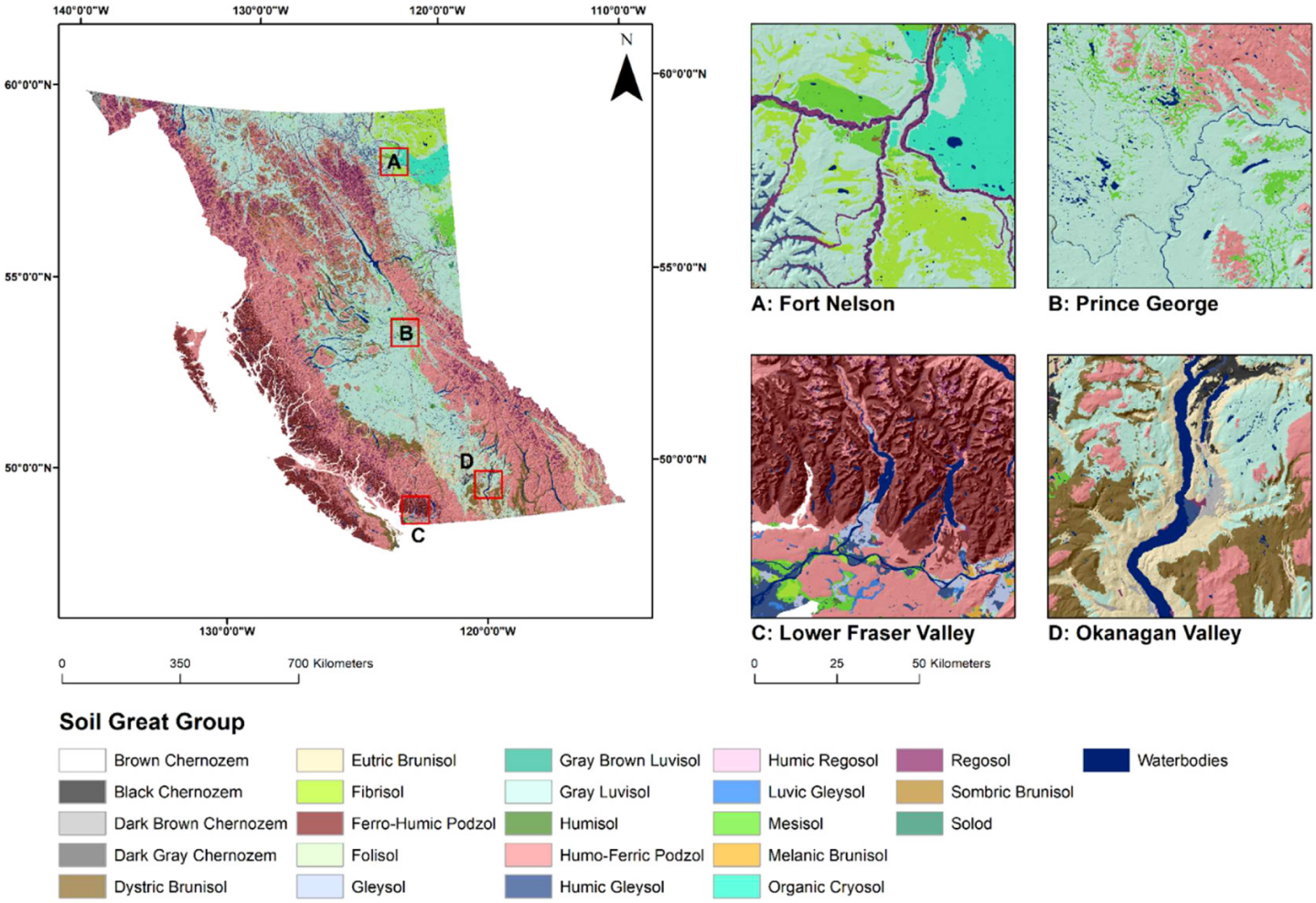


Table 3. Accuracy assessment of overall predictions, single-component map units, and predictions coinciding with single-component map units, using overall agreement (*C*), quantity disagreement (*Q*), and allocation disagreement (*A*).

	Overall					Single-component map units								
	Samples	Prediction accuracy				Samples	Survey accuracy				Prediction accuracy			
	n	κ	*C* (%)	*Q* (%)	*A* (%)	*n*	κ	*C* (%)	*Q* (%)	*A* (%)	κ	*C* (%)	*Q* (%)	*A* (%)
Parent material class	14 570	0.59	69	10	21	9856	0.53	64	15	21	0.63	74	9	17
Soil great group	14 316	0.37	55	27	17	9336	0.37	54	25	21	0.41	60	27	13
Soil order	14 316	0.41	62	22	16	9336	0.45	64	16	20	0.46	66	22	12

and may even develop into Humo-Ferric Podzols under mild and humid environments — especially in southern BC (Smith et al. 2011).

Overall, the soil types that developed through hydromorphic processes (i.e., Organic and Gleysol soils) were poorly predicted (Tables 5 and 6). The Luvic Gleysol, Humic Gleysol, and Gleysol great groups had accuracy rates of 16%, 13%, and 3%, respectively, and Fibrisol, Humisol, and Mesisol great groups had accuracy rates of 28%, 21%, and 8%, respectively. The Gleysol soils were most effectively mapped where the landscape was influenced by large-scale fluvial features — this is particularly true for the Lower Fraser Valley region of BC, where the soils were developed from the deltaic deposits that are found at low elevations due to the Fraser River (Fig. 3 inset C). However, the overall low accuracy rates of these hydromorphic soils may have been due to several factors. Firstly, the CSMs often included these soils as part of multicomponent map units because of their localized and often

Table 4. Confusion matrix between 14 570 observation points and predicted soil parent material using a random forest model.

	Predicted											
Actual	C	E	F	FG	L	LG	M	O	W	WG	Total	Accuracy rate (%)
C	**343**	0	17	27	1	1	137	0	45	0	571	60
E	1	**0**	0	2	0	1	4	0	0	0	8	0
F	54	0	**1682**	249	1	6	306	8	312	1	2619	64
FG	36	0	77	**787**	2	7	230	0	78	0	1217	65
L	1	0	33	27	**11**	28	44	0	1	0	145	8
LG	6	0	13	17	1	**25**	7	0	0	0	69	36
M	123	0	75	163	2	2	**2319**	3	364	2	3053	76
O	11	0	98	132	2	3	196	**178**	242	0	862	21
W	33	0	243	668	0	0	301	4	**4722**	1	5972	79
WG	1	0	0	11	0	0	4	1	26	**11**	54	20
Total	609	0	2238	2083	20	73	3548	194	5790	15		

Note: Bold values represent the diagonal of the confusion matrix and the number of correctly classified pixels for each class (see Table 2 for soil parent material codes).

limited occurrence in the landscape compared to other soils. Considering that the training data set we used only consisted of single-component map units, the representation of these soils was decreased. Furthermore, the map scale of the CSMs would have influenced the ability of the surveyor to delineate single-component instances of the hydromorphic soils at a 100 m spatial resolution. Lastly, an additional factor would also have been related to the 100 m spatial resolution, where the topographic variables derived from the 100 m DEM may not have highlighted the landscape features and hydromorphic processes with which these soils were associated. Each of these factors could have also contributed to the poor accuracy rates of Regosol soils due to their formation on highly disturbed sites (i.e., along floodplains and the bottom of steep slopes). Despite the large number of validation points ($n = 14\,316$), the Folisol, Organic Cryosol, and Solod great groups were not represented within the validation data set. Solod soils are rarely found in BC, while Folisol soils are typically found in association with Podzol soils. Organic Cryosols have a large distribution, regionally, in the Fort Nelson area of the province (Fig. 3 inset A); however, the validation data for that region were sparse (Fig. 1).

Variable importance

Variable importance plots were derived from the RF model outputs and were based on the MDG index, where variables with a high MDG value represented variables with greater importance (Fig. 4).

Soil parent material

Elevation, slope height, distance-to-nearest-stream, and topographic ruggedness index were identified as the four most important variables by the RF algorithm (Fig. 4). Elevation was particularly effective in distinguishing the parent materials that occur on mountainous and high-relief environments, where there was a clear separation of colluvial materials at the highest elevations and slope heights, which transitioned to till as the elevation decreased (Fig. 5). Elevation also plays a role in distinguishing fluvial and glacio-fluvial materials. Glacio-fluvial materials were typically mapped at lower elevations; however, the key distinguishing variables were the slope height and the distance-to-nearest-stream.

Soil taxonomic class

The variable importance plot for the prediction of soil great groups highlighted the strong influence of bioclimatic variables (Fig. 4). Mean annual precipitation was identified as the most important variable, followed by seven other bioclimatic variables; in contrast, the topographic variables that represented the local-scale variability of the environment were less important. Although we recognized elevation as a topographic covariate, it is inherently linked to climate due to the relationship between elevation and temperature via the environmental lapse rate. We believe that these variables were most effective in capturing the large-scale soil patterns of the province and the mapping of Brunisolic, Chernozemic, Luvisolic, and Podzolic soil orders, which had accuracy rates that ranged from 64% to 88%, and were markedly higher than the other soil orders, which had accuracy rates of <19%.

By inspecting the boxplots of the highest ranking climatic variables, based on the training data set, their relationship with the distribution of soil types was evident (Fig. 6). For example, Chernozemic and Podzolic soils, both recognized as "zonal soils", occupied distinct regions of covariate space, where Chernozemic soils were found in regions of the province with the lowest mean annual precipitation and the highest climatic moisture deficit, whereas the converse was true for Podzolic soils. The influence of climate on the distribution of the soil great groups was also captured in the training data, where increases in moisture (low values of climatic moisture deficit) and decreases in temperature (high values of degree days <0 °C) facilitated the transition of Brown Chernozems, to Dark Brown Chernozems, and to Black Chernozems as a result of the influence of climate on decomposition rates (Pennock et al. 2011). Similarly, the distinction between the great groups within the Podzolic or-

Table 5. Confusion matrix between 14 316 observation points and predicted soil great group using a random forest model.

	Predicted																								Accuracy rate
Actual	BC	BLC	DBC	DGC	DYB	EB	F	FO	FHP	G	GBL	GL	H	HFP	HG	HR	LG	M	MB	OC	R	SB	SO	Total	(%)
BC	**11**	3	1	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	46
BLC	0	**108**	11	1	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	125	86
DBC	0	3	**48**	7	4	22	0	0	0	0	0	7	0	0	0	0	0	0	0	0	1	0	0	92	52
DGC	0	5	1	**4**	0	4	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	18	22
DYB	0	0	1	0	**4226**	25	0	0	3	0	0	73	1	348	18	0	2	0	0	0	10	7	0	4714	90
EB	3	5	2	0	24	**186**	0	0	0	2	0	28	0	14	2	0	0	0	1	0	6	0	0	273	68
F	0	0	0	0	4	2	**5**	0	0	0	0	1	0	5	1	0	0	0	0	0	0	0	0	18	28
FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FHP	0	0	0	0	6	0	0	0	**73**	0	0	0	0	48	0	0	0	0	0	0	1	0	0	128	57
G	0	2	0	0	565	9	0	0	3	**24**	0	16	4	201	62	0	20	3	0	0	9	5	0	923	3
GBL	0	1	0	0	1	0	0	0	0	1	**1**	1	0	0	0	0	0	0	0	0	0	0	0	5	20
GL	0	3	0	0	29	20	0	0	0	2	0	**149**	1	13	0	0	0	1	0	0	4	0	0	222	67
H	0	0	0	0	189	0	0	0	1	0	0	7	**88**	114	12	0	2	6	0	0	3	0	0	422	21
HFP	0	0	0	0	1479	5	0	0	18	0	0	23	5	**2653**	13	0	13	7	0	0	3	7	0	4226	63
HG	1	24	1	0	793	3	1	0	2	4	0	22	10	306	**187**	0	35	2	0	0	2	14	0	1407	13
HR	0	2	1	0	109	4	0	0	0	0	0	2	2	23	7	**1**	0	0	0	0	3	0	0	154	1
LG	0	0	0	0	126	1	0	0	0	0	0	2	2	18	6	0	**32**	0	0	0	0	11	0	198	16
M	0	0	0	0	93	2	1	0	5	1	0	13	4	45	3	0	2	**14**	0	0	0	0	0	183	8
MB	1	2	2	2	4	19	0	0	0	4	0	8	0	4	2	0	1	0	**2**	0	4	0	0	55	4
OC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
R	0	2	0	0	194	11	0	0	19	2	0	9	1	55	13	0	1	2	0	0	**28**	0	0	337	8
SB	0	0	0	0	616	1	0	0	0	0	0	0	9	110	19	0	7	0	0	0	2	**28**	0	792	4
SO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
Total	16	160	68	14	8462	324	7	0	124	40	1	368	127	3957	346	1	115	35	3	0	76	72	0		

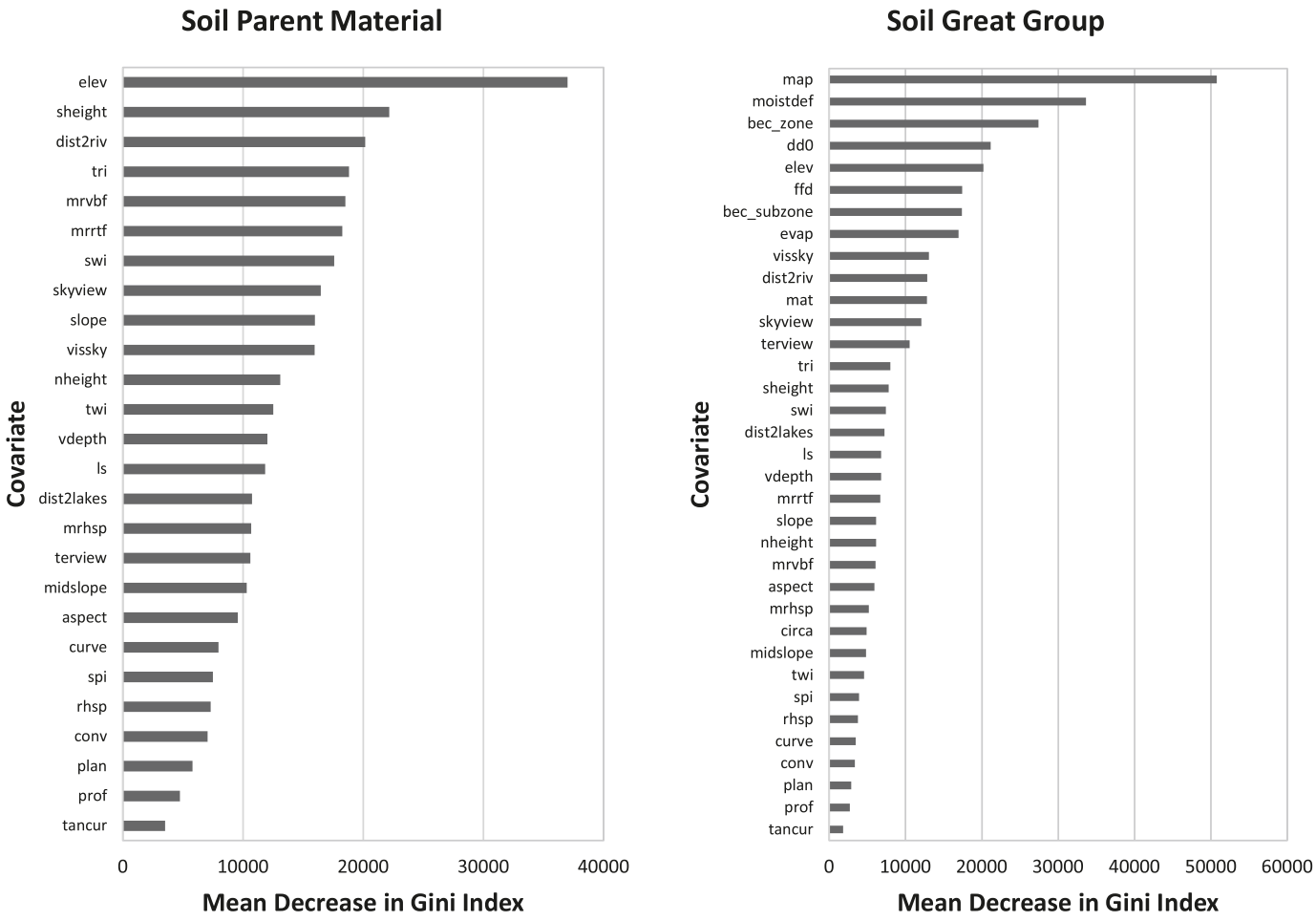
Note: Bold values represent the diagonal of the confusion matrix and the number of correctly classified pixels for each class (see Table 2 for great group codes).

Table 6. Confusion matrix between 14 316 observation points and predicted soil order using a random forest model.

	Predicted										Accuracy rate (%)
Actual	B	Ch	Cr	G	L	O	P	R	S	Total	
B	**5139**	18	0	57	109	10	479	22	0	5834	88
Ch	40	**203**	0	1	14	0	0	1	0	259	78
Cr	0	0	**0**	0	0	0	0	0	0	0	0
G	1527	28	0	**370**	40	22	530	11	0	2528	15
L	50	4	0	3	**151**	2	13	4	0	227	67
O	290	0	0	21	21	**118**	170	3	0	623	19
P	1497	0	0	26	23	12	**2792**	4	0	4354	64
R	318	5	0	23	11	5	97	**32**	0	491	7
S	0	0	0	0	0	0	0	0	**0**	0	0
Total	8861	258	0	501	369	169	4081	77	0		

Note: Bold values represent the diagonal of the confusion matrix and the number of correctly classified pixels for each class (see Table 2 for order codes).

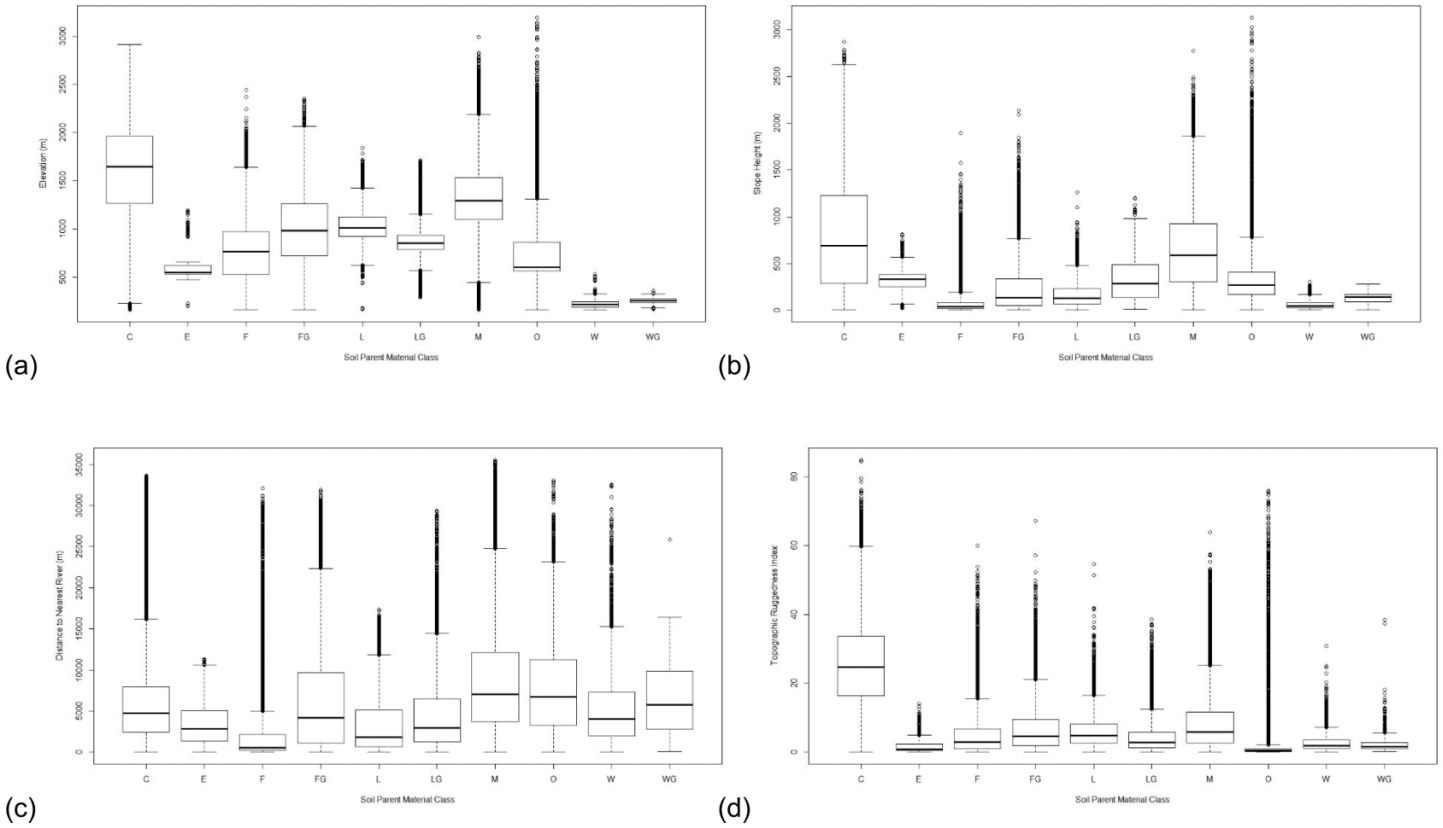
Fig. 4. Variable importance plots based on mean decrease in Gini index for soil parent materials and great group predictions using a random forest classifier. See Table 1 for the description of predictor variables.



der also appeared to be controlled by substantial differences in mean annual precipitation, where Ferro-Humic Podzols are located in wetter regions in comparison to Humo-Ferric Podzols.

Although the Brunisolic and Luvisolic soils, both considered to be "intrazonal" soil orders, were predicted at respectable accuracy rates, their separation in covariate space, in comparison to the Podzolic and Chernozemic soils, was

Fig. 5. Distribution of (*a*) elevation, (*b*) slope height, (*c*) distance-to-nearest-stream, and (*d*) topographic ruggedness index values based on training data and separated by soil parent material class (see Table 2 for parent material codes).



less clear. In particular, the distinction between Eutric and Dystric Brunisols, the two most common great groups from the Brunisolic order, was very subtle in that they were similar in terms of temperature range and topographic indices; however, their distributions were more obviously influenced by precipitation, where Dystric Brunisols (commonly associated with Podzolic soils) were located in moister regions of the province and at higher elevations.

In addition to the climatic variables, the vegetation indices were also determined to be important based on the RF algorithm (Fig. 4 and Table 7). Similar to the climatic variables, the "zonal soils" were strongly influenced by the vegetation patterns, where, for example, the transitions between Brown Chernozems to Dark Brown and Black Chernozems mirrored the transitions between BG to Ponderosa Pine and Interior Douglas-fir zones. In comparison, the Humo-Ferric Podzols were more closely associated with the Engelmann Spruce and Sub-Alpine zones, while the Ferro-Humic Podzols were predominantly restricted to the CWH zone — the wettest zone in the province. In comparison, the Canadian Land Cover Classification Circa 2000 data set was far less important than the BEC zone or subzone for predicting soil class. The information provided by an automated classification of individual pixels for land cover was influenced by short-range and temporal effects, such as natural disturbance, forest harvest, and agricultural practices, which are undoubtedly important factors affecting soil; however, the information was not specifically incorporated into the soil survey process and hence, not captured in the training data set.

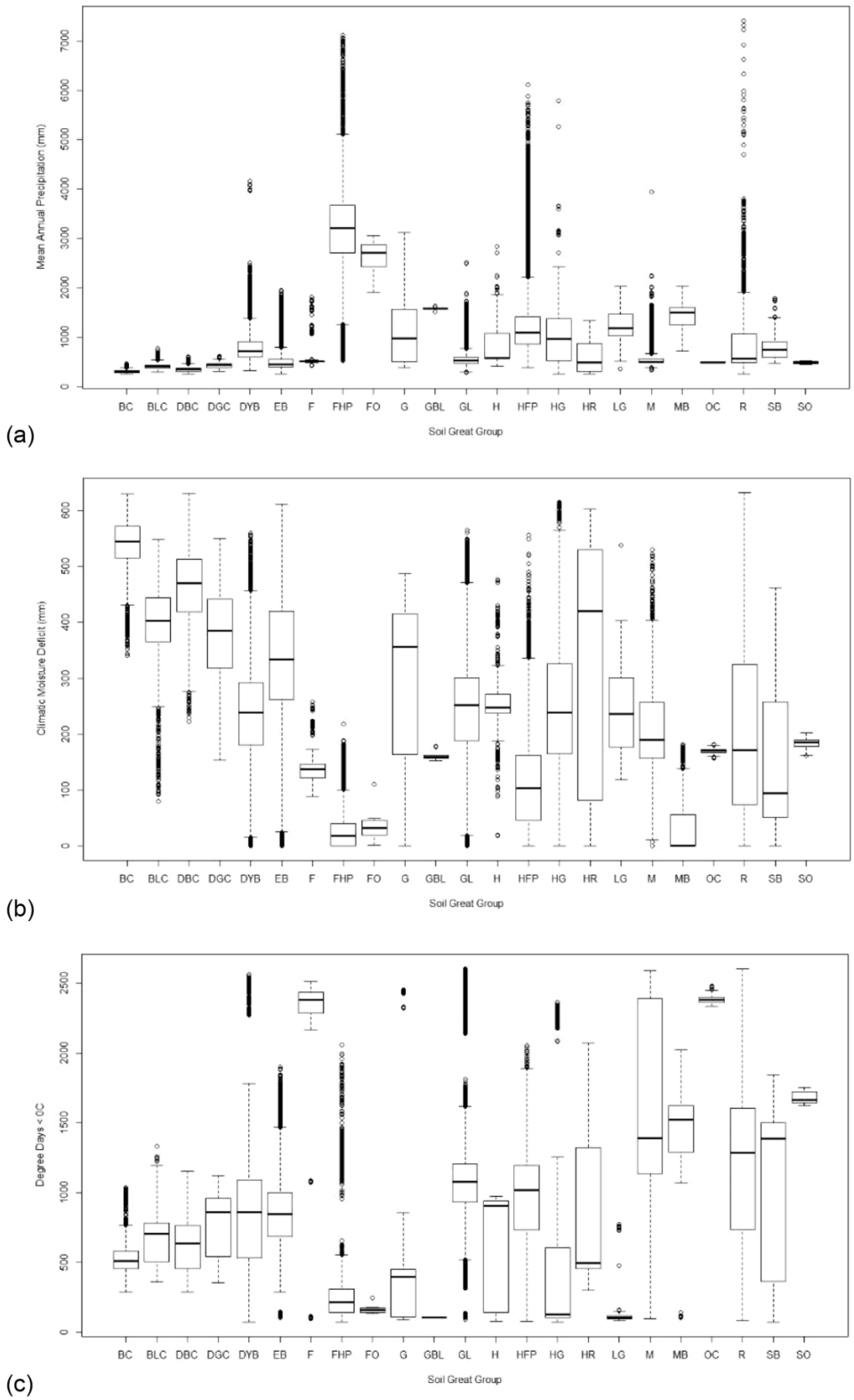
Prediction uncertainty

Class-probability surfaces for soil parent material (Fig. 7) and soil great groups (Fig. 8) were generated based on the 500 constituent decision trees of the RF algorithm. Spatial representations of ignorance uncertainty values are presented in Fig. 9; furthermore, Fig. 10 shows the ignorance uncertainty with respect to the validation points.

Soil parent material

Using the parent material probability surfaces, we were able to visualize the transitions in parent material, where fluvial materials were dominant in valleys with moderate to low slopes and moving toward higher slope positions, the dominant parent material transitioned to glacio-fluvial, to till, and to colluvium (Fig. 7). In cases where slopes of the valley bottom rapidly increased, the fluvial materials transitioned directly into colluvial materials. Overall, the combination of colluvium and till probability rasters was able to capture the major physiographic features of the province, where colluvial materials were most extensively mapped along the Coast, Columbia, Northern, and Rocky Mountains, and till was most extensively mapped along the Interior and Central Plateaus

Fig. 6. Distribution of (*a*) mean annual precipitation, (*b*) climatic moisture deficit, and (*c*) degree days <0 °C values based on training data and separated by soil great groups (see Table 2 for great group codes).



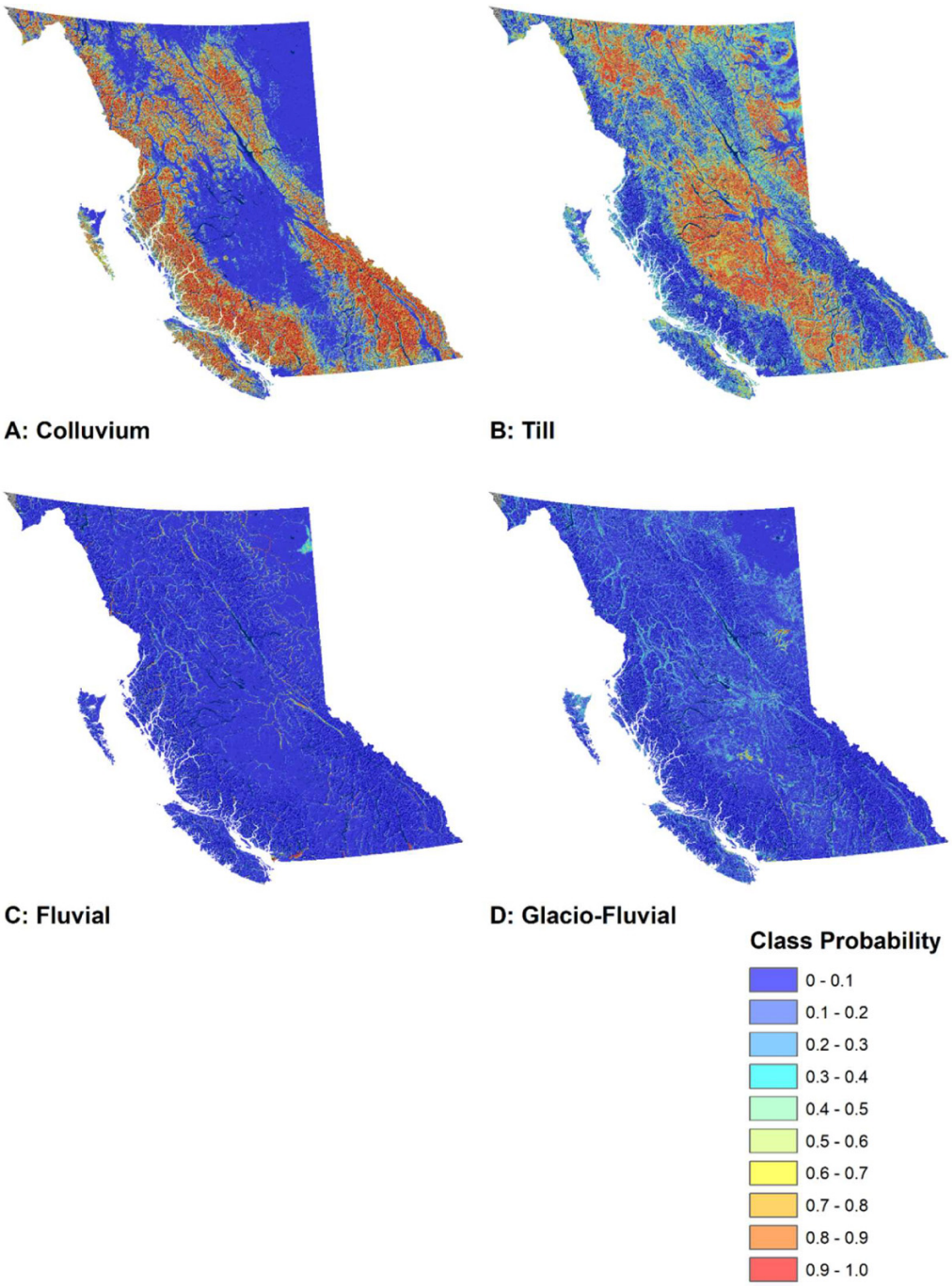
and The Great Plains regions of the province (Valentine et al. 1978).

The overall patterns of the prediction uncertainties, represented using the ignorance uncertainty measure, are shown in Fig. 9A. The ignorance uncertainty measure (*H*) had a mean of 0.33 with a standard deviation of 0.19. The spatial patterns of the ignorance uncertainty surface indicated that low-relief terrains resulted in the highest uncertainty

Table 7. Relative occurrence of training data points found under each biogeoclimatic ecosystem classification (BEC) zone and separated by soil great group (see Table 2 for great group codes).

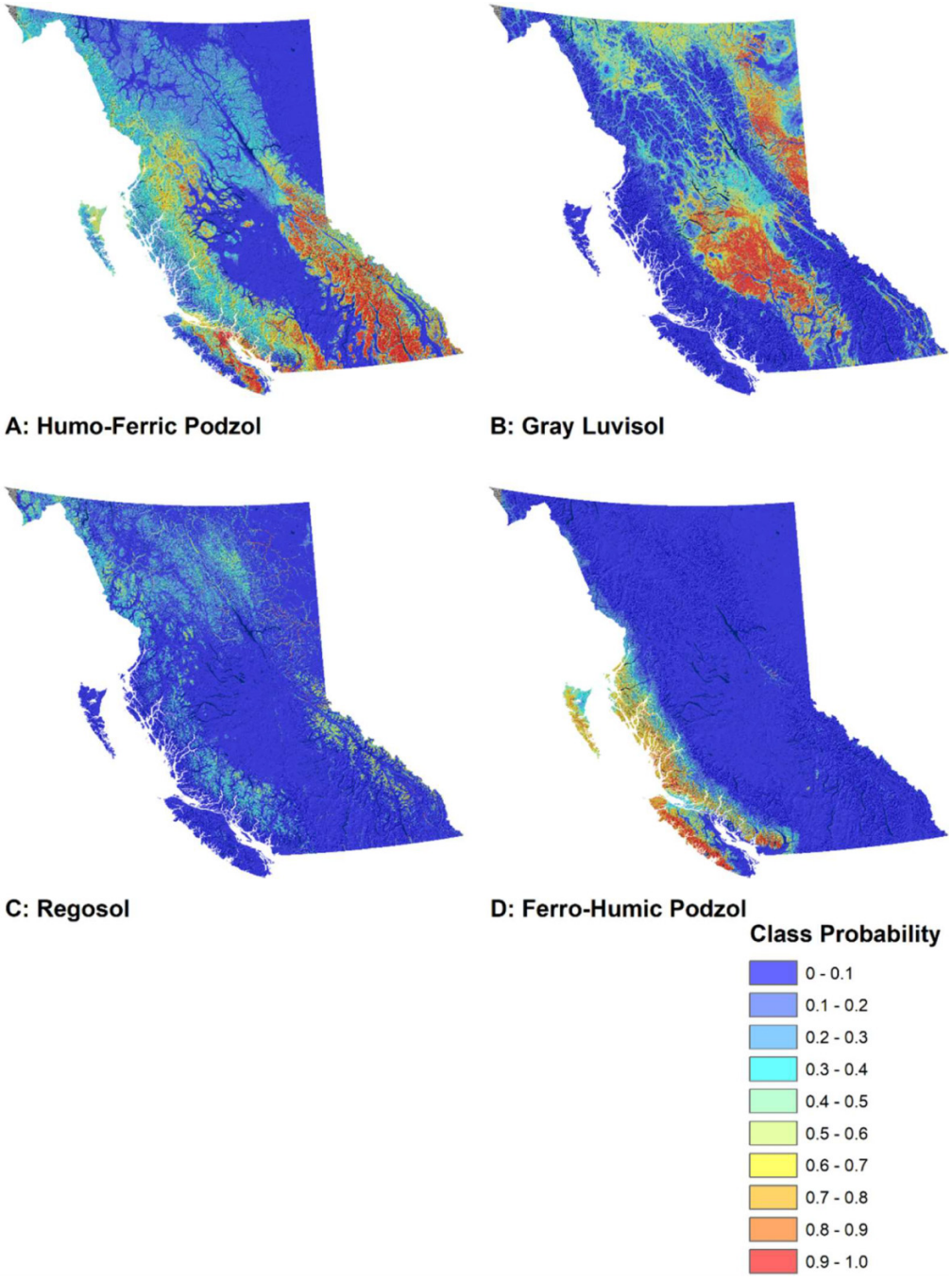
	By-class percentage of training data (%)																						
Biogeoclimatic zone	BC	BLC	DBC	DGC	DYB	EB	F	FHP	FO	G	GBL	GL	H	HFP	HG	HR	LG	M	MB	OC	R	SB	SO
Spruce–Willow–Birch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sub-Boreal–Spruce	0.0	0.0	0.0	3.7	7.9	0.9	0.0	1.1	0.0	1.9	0.0	15.6	14.9	1.5	2.3	0.0	0.0	16.6	1.0	0.0	7.1	0.0	0.0
Sub-Boreal Pine–Spruce	0.0	0.0	0.5	0.2	0.5	0.5	0.0	0.0	0.0	0.0	0.0	16.5	37.0	0.0	5.4	0.0	0.0	14.3	0.0	0.0	0.8	0.0	0.0
Ponderosa Pine	21.0	2.9	32.3	7.6	0.0	11.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	2.2	5.4	0.0	0.0	0.0	0.0	3.8	0.0	0.0
Mountain Spruce	0.0	1.6	0.0	0.1	20.3	15.4	0.2	0.1	0.0	0.0	0.0	12.5	0.0	2.7	1.5	0.0	0.0	6.1	0.0	0.0	5.1	1.3	0.0
Mountain Hemlock	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.2	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
Interior Mountain-heather Alpine	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.0	1.6	11.4	0.0
Interior Douglas-fir	6.9	85.4	32.6	75.7	7.8	56.7	0.0	0.0	0.0	7.6	0.0	29.9	12.7	0.3	15.9	12.4	2.1	5.4	0.0	0.0	12.2	0.9	0.0
Interior Cedar–Hemlock	0.0	0.0	0.0	0.0	37.4	4.7	0.0	0.0	0.0	27.7	0.0	7.2	3.5	24.4	1.5	16.7	0.0	0.4	0.3	0.0	9.4	19.6	0.0
Engelmann Spruce–Subalpine Fir	0.0	0.8	0.0	0.0	14.6	8.1	0.0	3.9	0.0	0.0	0.0	4.9	0.0	53.1	1.1	24.2	0.0	2.7	64.4	0.0	19.1	49.1	0.0
Coastal Western Hemlock	0.0	0.0	0.0	0.0	5.8	0.2	1.1	87.5	100.0	57.7	100.0	0.1	21.3	16.4	30.1	0.0	47.6	2.7	21.0	0.0	4.3	3.4	0.0
Coastal Mountain-heather Alpine	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Coastal Douglas-fir	0.0	0.0	0.0	0.0	4.9	0.0	1.9	0.0	0.0	2.4	0.0	0.0	10.6	0.6	27.5	0.0	50.4	1.7	0.0	0.0	0.2	3.7	0.0
Boreal White and Black Spruce	0.0	0.0	0.0	0.0	0.7	1.5	96.9	0.0	0.0	2.6	0.0	13.1	0.0	0.0	9.9	0.0	0.0	50.0	0.0	100.0	29.1	0.0	100.0
Bunchgrass	72.1	9.3	34.6	12.6	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	36.6	0.0	0.0	0.0	0.0	2.4	0.0	0.0
Boreal Altai Fescue Alpine	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	4.8	0.0	0.0	12.9	0.0	4.6	10.6	0.0

Fig. 7. Class-probability surfaces based on 500 decision trees of the random forest model at a 100 m spatial resolution for BC. Most frequently occurring parent material classes include colluvium (A), till (B), fluvial (C), and glacio-fluvial (D) materials (BC Albers projection). [Colour online]



(i.e., $H > 0.70$). Low-lying areas of coastal BC were particularly challenging to predict due to a combination of complex geomorphic processes, which resulted in the deposition of fluvial, glacio-fluvial, and till materials — all of which may occupy similar regions of variables space (Fig. 5); furthermore, the presence of marine and glacio-marine materials also complicated the prediction. Regions of the province that consist of wide valleys and terraced features were also challenging to predict due to the presence of fluvial, glacio-fluvial, till, and glacio-lacustrine materials. An example of such a landscape includes the Rocky Mountain Trench, where a similar combination of parent materials has been identified in Clague (1975). Lastly, the combination of organic and fluvial parent materials, east of the Fort Nelson region, resulted in high values of ignorance uncertainty.

Fig. 8. Class-probability surfaces based on 500 decision trees of the random forest model at a 100 m spatial resolution for BC. Most frequently occurring great groups include Humo-Ferric Podzols (A), Gray Luvisols (B), Regosols (C), and Ferro-Humic Podzols (D) (BC Albers projection). [Colour online]

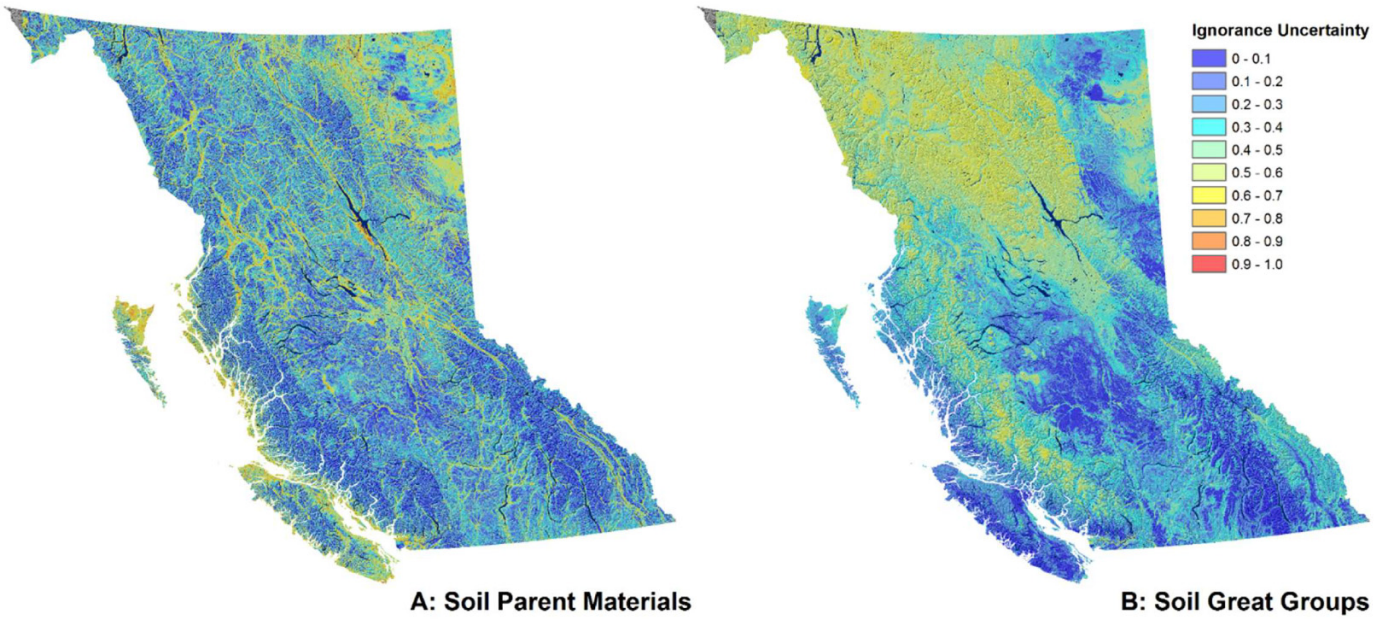


Soil taxonomic class

Large-scale transitions in soil great groups were clearly visible using the probability surfaces. The influence of orographic precipitation was most clear for the mountain slopes of Coastal BC, where the presence of Ferro-Humic Podzols was most dominant along the windward side of the Coast Mountains, and proceeding eastward, there was an abrupt transition to Humo-Ferric Podzols (Fig. 8). In contrast, as the drier air moves across the Interior Plateau and then up and over the Columbia Mountains, the resulting orographic influence produces less precipitation than along the Coast; hence, the development of Ferro-Humic Podzols is limited, while Humo-Ferric Podzols become more common.

For large portions of BC's Interior Plateau, where there is less precipitation, Gray Luvisolic soils dominate the predictions. It was of interest that the areas with high probabilities of being predicted as Gray Luvisolic soils also coincided with the prediction of till as parent material. The main process involved with the formation of Luvisolic soils, lessivage,

Fig. 9. Ignorance uncertainty surfaces for soil parent material (A) and soil great group (B) predictions based on a random forest model at a 100 m spatial resolution for BC (BC Albers projection). [Colour online]



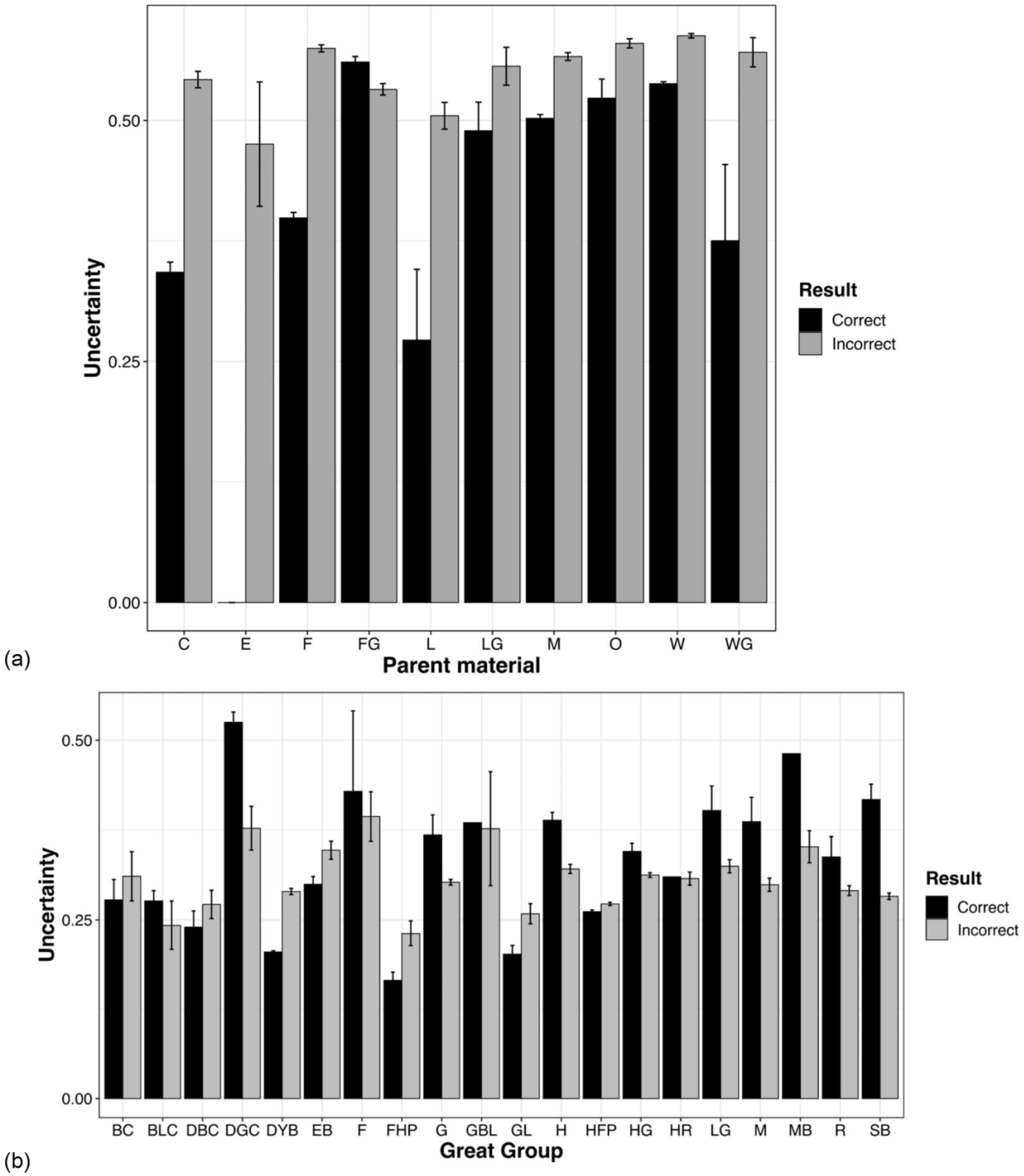
depends on the initial presence of clay particles within the glacial sediments. Although we did not use geology as an independent input for our predictions, it is well known in BC that the bedrock of the Coast Mountains has a very high component of acidic intrusive rocks that are expected to weather to sand during pedogenesis. In contrast, the Interior Plateau is made up of a more diverse assemblage of rock types, including volcanic, sedimentary, and metamorphic rocks that produce medium- and fine-textured surficial materials upon weathering. Given that the parent material predictions were derived only using topographic indices, we suggest that the large-scale topographic characteristics of the province (i.e., the transition from Coast Mountains to Interior Plateau) coincide not only with reduced precipitation, but also with the presence of more clay in the soil parent materials. Therefore, two of the most important characteristics of the pedogenic environment distinguishing Luvisols from Podzols occur to the east of this transition, but not to the west.

In summary, high precipitation levels in the western parts of the Coast Mountains favour the development of Podzols over Luvisols, while medium–fine textured soils on the Plateau to the east of the Coast Mountains favour the development of Luvisols over Brunisols. The Regosolic great group was the third-most extensively predicted soil class for the province; however, their distributions were mainly predicted based on topography and tied to specific geographic features. In southern BC, the highest probability values were observed at high elevations with steep slopes along the Coast and Columbia Mountains, where erosion processes inhibit pedogenic processes and also within valley bottoms, where fluvial materials are deposited. However, at high latitudes, Regosolic soils were commonly associated with the presence of cirque — as a result of ongoing alpine glaciation.

The ignorance uncertainty, H, for the soil great group had a mean of 0.36 with a standard deviation of 0.17 (Fig. 10*b*). Unlike the parent material predictions, where high uncertainty values were found in closer association with landscape features throughout the province, the areas of high uncertainty values for the soil great group map were primarily concentrated in regions of the province where CSMs were not available — especially the northwestern region of the province. A second observation was that the northern portions of BC also experienced low mean annual temperatures and areas that had moderate to high uncertainty values (i.e., $H > 0.5$) coincided with mean annual temperatures of <-3 °C. As a result, it is likely that the higher uncertainty values for the region were due to the lack of training data that represented the coldest part of the province. In addition, pixels where $H > 0.6$ also appeared to be related to topography and slope position for the (sub-)alpine landscapes, where the greatest uncertainty values were observed along lower backslopes. Although these regions were typically classified as Humo-Ferric Podzols, the probability values for Gray Luvisols, Humo-Ferric Podzols, Regosols, and Sombric Brunisols were similar and often found in association with each other (Valentine et al. 1978; Trowbridge 1994).

Incorrect predictions for soil taxonomic class were often associated with low ignorance uncertainty values (Fig. 10*b*), in contrast to soil parent materials. The lowest ignorance uncertainty values were associated with correct prediction for Dystric Brunisols, Ferro-Humic Podzols, and Gray Luvisols. Ferro-Humic Podzols had a distinct feature space for mean annual precipitation, but this was not evident for Dystric Brunisols or Gray Luvisols. Dystric Brunisols were well represented in the training data set, accounting for 31% of all points. The

Fig. 10. Ignorance uncertainty and prediction accuracy for soil parent material (*a*) and soil great group (*b*) predictions based on a random forest model at a 100 m spatial resolution for BC.



highest ignorance uncertainty was for the correct predictions of Dark Brown Chernozems.

For the parent material validation points with correct predictions, ignorance uncertainty values were lower, except for eolian and fluvial materials (Fig. 10*a*). The lowest ignorance uncertainty values were obtained for correct predictions of colluvium, lacustrine, and glacio-marine. Two of these parent material classes (colluvium and glacio-marine) had a distinct feature space for at least one predictor in Fig. 5, and they were all relatively sparse in the training data set, with a combined representation of less than 5%.

General discussion

Even though there was a lack of training and validation data across large regions of the province, and especially Northwestern BC, the resulting maps were consistent with

our pedological understanding of the environment. In terms of the parent material map, one of the surprising outcomes of the map was in the prediction of glacio-marine for the northeastern part of Haida Gwaii, where training data for that parent material class were derived exclusively from the Lower Fraser Valley in southwestern BC. Although there were no validation points located on Haida Gwaii, the extrapolation of that parent material class was consistent with a previous geological reconnaissance study for the same region (Clague et al. 1982). In terms of the prediction of soil great groups for the (sub-)alpine regions of Northwestern BC, the predicted soils were also consistent with their spatial patterns as described in the literature (Valentine et al. 1978).

This study used the BCSIS data set to validate the model predictions and assumed that the soil parent material and taxonomic units were correctly identified in the data set. However, such assumptions were also a limiting factor and a source of uncertainty in this study because there were many instances in the BCSIS data set where there were no corresponding analytical measurements to support the decisions made by the surveyors. For example, the distinction between Dystric Brunisols and Eutric Brunisols is based on a pH threshold of 5.5 using 0.01 mol/L $CaCl_2$ for the B horizon, whereby the pH may or may not have been measured. A similar uncertainty may be introduced in identifying soil parent materials as well due to inconsistencies in how till and colluvium or fluvial and glacio-fluvial materials are differentiated by different surveyors. Furthermore, it should be recognized that there were also inconsistencies between the detailed soil surveys whereby the glacio-marine materials were not differentiated from the marine materials in the southern Vancouver Island surveys. Lastly, due to the high relief of BC, we also recognize that validation points were most likely to be located in agriculturally intensive regions and where access was good. Hence, this could possibly explain the underrepresentation of colluvium observations where accessibility may be limited and the potential overrepresentation of fluvial observations is found in agricultural regions (i.e., Lower Fraser Valley).

To improve upon this study, alternative mapping approaches that are focused on the prediction of "azonal" soils (e.g., Organics, Gleysols, and Regosols) could be tested. For example, Bulmer et al. (2016) used data from a wetland inventory of BC to simply assign, or "burn in" organic soils to known wetland locations. The amount of information contained in our training data was very limited for Gleysols and Regosols because either their occurrence was so localized that they were simply not mapped by surveyors, given the mapping scale, or they were recognized as part of a multicomponent map unit. Another issue related to the spatial resolution of the topographic indices was that the effective mapping of the azonal soils may require a finer resolution DEM, which would be more effective in capturing the small-scale topographic variability and be more reflective of the localized, pedogenic processes for these soils. A potential solution may be to use a suite of topographic indices derived at multiple spatial resolutions (Smith et al. 2006; Behrens et al. 2010) or the application of wavelet transformation to spatially decompose the topographic indices (Taghizadeh-Mehrjardi et al. 2021); however, such an approach would still require a DEM with finer resolution than the one that was available for this study.

Conclusions

This study aimed to extend the existing DSM approaches that were previously tested in BC for the development of provincial-scale maps of soil taxonomic classes and soil parent material classes using the RF classification algorithm. The key findings are summarized as follows:

- By training a machine learner using single-component map units from a CSM, the resulting DSM had improved accuracy rates. The DSM approach increased accuracy rates by 10%, 6%, and 2% when mapping soil parent material classes, soil great group, and soil order, respectively. These results were consistent with similar studies such as Collard et al. (2014) and Yang et al. (2011) and reinforce the findings of Heung et al. (2017).
- When mapping soil taxonomic classes, it was observed that zonal (Podzols and Chernozems) and intrazonal soils (Brunisols and Luvisols) were far more effectively mapped in comparison to azonal soils (Gleysols, Regosols, and Organics). We believe that the ineffective mapping of azonal soils was, in part, due to a combination of the spatial resolution of the environmental variables and the mapping scale of the CSMs from which the training data were derived.
- Through the examination of variable importance plots, it was observed that the predicted distribution of soil taxonomic classes was heavily influenced by climatic variables and, to a lesser degree, land cover data. Based on the covariate boxplots, it was observed that mean annual precipitation was particularly important in distinguishing the Podzolic, Chernozemic, and Luvisolic soil orders.

Overall, this study, in conjunction with Heung et al. (2014, 2016, 2017) and Bulmer et al. (2016), demonstrates the long-term value of CSMs in BC and their potential to be refined using DSM and machine-learning techniques. The overall framework of training models using legacy maps may have the potential to be adopted for the refinement of legacy ecological or geological maps that share a similar data structure to CSMs.

Acknowledgement

The authors would like to thank the hard work and dedication of soil surveyors in British Columbia that left a legacy of maps that formed the basis of this study.

Article information

History dates

Received: 19 July 2021
Accepted: 4 January 2022
Accepted manuscript online: 4 February 2022
Version of record online: 29 August 2022

Notes

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

Copyright

Author information

Author notes

Authors Brandon Heung and Chuck E. Bulmer served as Guest Editors at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by Angela Bedard-Haughn.

Author contributions

B.H., conceptualization, data curation, formal analysis, methodology, validation, visualization, and writing (original draft); C.B., formal analysis, methodology, and writing (original draft, review, and editing); M.S. and J.Z., writing (original draft, review, and editing).

Competing interests

The authors declare that no competing interests exist.

Funding information

The authors declare no specific funding for this work.

References

Agriculture and Agri-Food Canada. 2021. Canadian soil information service [online]. Availablefrom https://sis.agr.gc.ca/cansis/[accessed 22 June 2016].

Anderson, D.W., and Smith, C.A.S. 2011. A history of soil classification and soil survey in Canada: personal perspective. Can J. Soil Sci. **91**: 675–695. doi:10.4141/cjss10063.

B.C. Ministry of Agriculture and B.C. Ministry of Environment. 2018. British Columbia soil information finder tool [online]. Availablefrom https://governmentofbc.maps.arcgis.com/apps/MapSeries/index.html?appid=cc25e43525c5471ca7b13d639bbcd7aa.

B.C. Ministry of Sustainable Resource Management. 2002. Gridded digital elevation model product specification[online]. 2nd ed. Digital Elevation Model, B.C. Ministry of Sustainable Resource Management, British Columbia. Availablefrom https://www2.gov.bc.ca/gov/content/data/geographic-data-services/topographic-data/elevation/digital-elevation-model.

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. doi:10.1016/j.geoderma.2009.07.010.

Bertrand, R.A., Hughes-Games, G.A., and Nikkel, D.C. 1991. Soil management handbook for the Lower Fraser Valley. 2nd ed. B.C. Ministry of Agriculture, Fisheries and Food, Abbotsford, BC.

Blackford, C., Heung, B., Baldwin, K., Fleming, R.L., Hazlett, P.W., Morris, D.M., et al. 2021. Digital soil mapping workflow for forest resource applications: a case study in the hearst forest Ontario. Can. J. For. Res. **51:** 59–77. doi:/cjfr-2020-0066. PMID: 10.1139.

Breiman, L. 2001. Random forests. Mach. Learn. **45**: 5–32. doi:10.1023/A:1010933404324.

Brown, G.W. 1973. The impact of timber harvest on soil and water resources. Bull. 827. Oregon State University, Extension Service, Corvallis, OR.

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. 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, Singapore. pp. 291–303. doi:10.1007/978-981-10-0415-5_24.

Carré, F., McBratney, A.B., Mayr, T., and Montanarella, L. 2007. Digital soil assessments: beyond DSM. Geoderma, **142**: 69–79. doi:10.1016/j.geoderma.2007.08.015.

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

Church, M., and Ryder, J.M. 2010. Physiography of British Columbia [online]. In Compendium of forest hydrology and geomorphology in British Columbia. B.C. Ministry of Forests and Range, For. Sci. Prog., Victoria, BC. p. 66. Available from https://www.for.gov.bc.ca/hfd/pubs/docs/lmh/Lmh66.htm.

Clague, J.J. 1975. Late quaternary sediments and geomorphic history of the southern Rocky Mountain Trench, British Columbia. Can. J. Earth Sci. **12**: 595–605. doi:10.1139/e75-054.

Clague, J., Mathewes, R., and Warner, B. 1982. Late quaternary geology of eastern Graham Island, Queen Charlotte Islands, British Columbia. Can. J. Earth Sci. **19**: 1786–1795. doi:10.1139/e82-157.

Collard, F., Kempen, B., Heuvelink, G.B.M., Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., et al. 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Reg. **1**: 21–30. doi:10.1016/j.geodrs.2014.07.001.

DeLong, C., Annas, R., and Stewart, A. 1991. Boreal White and Black Spruce zone. In Ecosystems of British Columbia. Research Branch Ministry of Forests.

Fenger, M.A., and Kowall, R.C. 1992. Biophysical soil landscapes inventory of the Stikine–Iskut area (Mapsheets 104F, 104 G, and Parts of 104B and 104H). Soil Survey Report, B.C. Ministry of Environment, Land and Parks.

Geng, X., Fraser, W., VandenBygaart, B., Smith, S., Waddell, A., Jiao, Y., et al. 2010. Toward digital soil mapping in Canada: existing soil survey data and related expert knowledge. *In* Digital soil mapping: bridging research, environmental application, and operation. Edited by J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink and S. Kienast-Brown. Springer Netherlands, Dordrecht. pp. 325–335. doi:10.1007/978-90-481-8863-5_26.

Goodchild, M.F., Chih-Chang, L., and Leung, Y. 1994. Visualizing fuzzy maps. Visualization in geographical information systems. John Wiley & Sons, New York. pp. 158–167.

Hartemink, A.E., and McBratney, A.B. 2008. A soil science renaissance. Geoderma, **148**: 123–129. doi:10.1016/j.geoderma.2008.10.006.

HectaresBC. 2012. Hectares BC [online]. Availablefrom https://hectaresbc.ca/app/habc/HaBC.html[accessed 28 October 2016].

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., et al. 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One, **12**: e0169748. doi:10.1371/journal.pone.0169748. PMID: 28207752.

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., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C., and Schmidt, M. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, **265**: 62–77. doi:10.1016/j.geoderma.2015.11.014.

Heung, B., Hodúl, M., and Schmidt, M. 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for

mapping soil classes. Geoderma, **290**. doi:10.1016/j.geoderma.2016.12.001.

Holland, S.S. 1976. Landforms of British Columbia: a physiographic outline[online]. Department of Mines and Petroleum Resources. Availablefrom https://books.google.ca/books?id=jGlhuwEACAAJ.

Integrated Land Management Bureau. 2010. Freshwater Water Atlas User Guide[online]. Integrated Land Management Bureau, Victoria, BC. Availablefrom https://www2.gov.bc.ca.

Kelley, C.C., and Spilsbury, R.H. 1939. Soil survey of the Lower Fraser Valley. Technical bulletin 20. British Columbia Department of Agriculture Co-Operating with Experimental Farms Service. Dominion Department of Agriculture, Ottawa, ON.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., and Stoorvogel, J.J. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma, **151**: 311–326. doi:10.1016/j.geoderma.2009.04.023.

Kuhn, M. 2008. Building predictive models in R using the caret package [online]. Available from https://www.jstatsoft.org/v028/i05.

Lal, R. 2004. Soil carbon sequestration impacts on global climate change and food security. Science, **304**: 1623. doi:10.1126/science.1097396. PMID: 15192216.

Leung, Y., Goodchild, M.F., and Lin, C.-C. 1993. Visualization of fuzzy scenes and probability fields. Comput. Sci. Stat. 416–416.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. R News, **2**: 18–22.

Luttmerding, H.A. 1981. Soils of the Langley-Vancouver map area. Report No. 15, British Columbia Soil Survey. BC Ministry of Environment, Kelowna, BC.

MacMillan, R.A., Moon, D.E., and Coupé, R.A. 2007. Automated predictive ecological mapping in a forest region of BC, Canada, 2001–2005. Geoderma, **140**: 353–373. doi:10.1016/j.geoderma.2007.04.027.

MacMillan, R.A., Moon, D.E., Coupé, R.A., and Phillips, N. 2010. Predictive ecosystem mapping (PEM) for 8.2 million ha of forestland, British Columbia, Canada. *In* Digital soil mapping. Springer. pp. 337–356.

McKeague, J.A., and Stobbe, P.C. 1978. History of soil survey in Canada 1914–1975. Canada Department of Agriculture.

Meidinger, D., and Pojar, J. 1991. Ecosystems of British Columbia. Special report series. Ministry of Forests, British Columbia.

Monger, J.W.H. 1997. Plate tectonics and northern Cordilleran geology: an unfinished revolution. Geoscience Canada.

Nicholson, A., Hamilton, E., Harper, W.L., and Wikeem, B.M. 1991. Bunchgrass zone. In Ecosystems of British Columbia. Edited by D. Meidinger and J. Pojar. Ministry of Forests, Victoria, BC. pp. 125–137.

Olthof, I., Latifovic, R., and Pouliot, D. 2009. Development of a circa 2000 land cover map of northern Canada at 30 m resolution from Landsat. Can. J. Remote Sens. **35**: 152–165. doi:10.5589/m09-007.

Pennock, D., Bedard-Haughn, A., and Viaud, V. 2011. Chernozemic soils of Canada: genesis, distribution, and classification. Can. J. Soil Sci. **91**: 719–747. doi:10.4141/cjss10022.

Pojar, J., Klinka, K., and Dermarchi, D.A. 1991. Coastal Western Hemlock zone. In Ecosystems of British Columbia. Ministry of Forests Victoria, BC. pp. 95–111.

Pontius, R., and Millones, M. 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. **32**: 4407–4429. doi:10.1080/01431161.2011.552923.

R Development Core Team. 2012. R: a language and environment for statistical computing[online]. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.r-project.org/[accessed 26 February 2016].

SAGA Development Team. 2011. System for automated geoscientific analyses (Version 2.0. 6.)[online]. SAGA User Group Association, Hamburg, Germany. Available from www.saga-gis.org[accessed 14 September 2013].

Scarpone, C., Schmidt, M.G., Bulmer, C.E., and Knudby, A. 2016. Modelling soil thickness in the critical zone for southern British Columbia. Geoderma, **282**: 59–69. doi:10.1016/j.geoderma.2016.07.012.

Scarpone, C., Schmidt, M.G., Bulmer, C.E., and Knudby, A. 2017. Semi-automated classification of exposed bedrock cover in British Columbia's southern mountains using a random forest approach. Geomorphology, **285**: 214–224. doi:10.1016/j.geomorph.2017.02.013.

Schut, P., Smith, S., Fraser, W., Geng, X., and Kroetsch, D. 2011. Soil Landscapes of Canada: building a national framework for environmental information. Geomatica, **65**: 293–309. doi:10.5623/cig2011-045.

Shi, X. 2010. ArcSIE user's guide. SIE LLC, NH, USA.

Shi, X., Long, R., Dekett, R., and Philippe, J. 2009. Integrating different types of knowledge for digital soil mapping. Soil Sci. Soc. Am. J. **73**: 1682–1692. doi:10.2136/sssaj2007.0158.

Smith, M.P., Zhu, A.-X., Burt, J.E., and Stiles, C. 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma, **137**: 58–69. doi:10.1016/j.geoderma.2006.07.002.

Smith, C.A.S., Webb, K.T., Kenney, E., Anderson, A., and Kroetsch, D. 2011. Brunisolic soils of Canada: genesis, distribution, and classification. Can. J. Soil. Sci. **91**: 695–717. doi:10.4141/cjss10058.

Smith, C.A.S., Daneshfar, B., Frank, G., Flager, E., and Bulmer, C. 2012. Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. *In* Digital soil assessments and beyond. Edited by B. Minasny, B.P. Malone and A.B. McBratney. CRC Taylor & Francis.

Smith, S., Neilsen, D., Frank, G., Flager, E., Daneshfar, B., Lelyk, G., et al. 2016. Disaggregation of legacy soil maps to produce a digital soil attribute map for the Okanagan Basin, British Columbia, Canada. In Digital soil mapping across paradigms, scales and boundaries. Springer. pp. 305–317.

Soil Classification Working Group, Canadian Agricultural Services Coordinating Committee Soil Classification Working Group, National Research Council Canada, Canada Agriculture, and Agri-Food Canada Research Branch. 1998. The Canadian system of soil classification. 3rd ed. NRC Research Press.

Sondheim, M., and Suttie, K. 1983. User manual for the British Columbia soil information system. Ministry of Forests, Province of British Columbia.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., and Triantafilis, J. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region. Iran. Geoderma, **253**: 67–77. doi:10.1016/j.geoderma.2015.04.008.

Taghizadeh-Mehrjardi, R., Schmidt, K., Toomanian, N., Heung, B., Behrens, T. Mosavi, A., et al. 2021. Improving the spatial prediction of soil salinity in arid regions using wavelet transformation and support vector regression models. Geoderma, **383**: 114793. doi:10.1016/j.geoderma.2020.114793.

Teague, W.R., Dowhower, S.L., Baker, S.A., Haile, N., DeLaune, P.B., and Conover, D.M. 2011. Grazing management impacts on vegetation, soil biota and soil chemical, physical and hydrological properties in tall grass prairie. Agric. Ecosyst. Environ. **141**: 310–322. doi:10.1016/j.agee.2011.03.009.

Trowbridge, R. 1994. Field survey of the Tatshenshin/Alsek region: soil reconnaissance report. BC Ministry of Forests, Smithers, BC.

Valentine, K.W.G. 1986. Soil resource surveys for forestry. Soil, terrain, and site mapping in boreal and temperate forests. Clarendon Press, Oxford University Press, NY.

Valentine, K.W.G., Sprout, P.N., Baker, T.E., and Lavkulich, L.M. 1978. The soil landscapes of British Columbia[online]. Resource Analysis Branch, Ministry of the Environment, Victoria, BC. Available from: www.env.gov.bc.ca/soils/landscape/index.html.

Wang, T., Hamann, A., Spittlehouse, D.L., and Murdock, T.Q. 2012. ClimateWNA — high-resolution spatial climate data for western North America. J. Appl. Meteorol. Climatol. **51**: 16–29. doi:10.1175/JAMC-D-11-043.1.

Warrens, M.J. 2015. Properties of the quantity disagreement and the allocation disagreement. Int. J. Remote Sens. **36**: 1439–1446. doi:10.1080/01431161.2015.1011794.

Wilding, L.P., and Lin, H. 2006. Advancing the frontiers of soil science towards a geoscience. Geoderma, **131**: 257–274. doi:10.1016/j.geoderma.2005.03.028.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.-X., Hann, S., Burt, J.E., et al. 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. **75**: 1044–1053. doi:10.2136/sssaj2010.0002.

Zhu, A.-X. 1997. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Photogramm. Eng. Remote Sens. **63**: 1195–1201.