Geoderma 235–236 (2014) 59–73

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the *k*-nearest neighbor method

Nicolas Mansuy *, Evelyne Thiffault, David Paré, Pierre Bernier, Luc Guindon, Philippe Villemaire, Vincent Poirier, André Beaudoin

Natural Resources Canada, Canadian Forest Service, P.O. Box 10380, Stn. Ste-Foy, Québec, QC G1V 4C7, Canada

ARTICLE INFO

Article history:
Received 3 March 2014
Received in revised form 5 June 2014
Accepted 30 June 2014
Available online xxxx

Keywords:
Climate
Raster
SCORPAN
Soil texture
Topography

ABSTRACT

Large-scale mapping of soil properties is increasingly important for environmental resource management. While forested areas play critical environmental roles at local and global scales, forest soil maps are typically at low resolution. The objective of this study was to generate continuous national maps of selected soil variables (C, N and soil texture) for the Canadian managed forest landbase at 250 m resolution. We produced these maps using the *k*NN method with a training dataset of 538 ground-plots from the National Forest Inventory (NFI) across Canada, and 18 environmental predictor variables. The best predictor variables were selected (7 topographic and 5 climatic variables) using the Least Absolute Shrinkage and Selection Operator method. On average, for all soil variables, topographic predictors explained 37% of the total variance versus 64% for the climatic predictors. The relative root mean square error (RMSE%) calculated with the leave-one-out cross-validation method gave values ranging between 22% and 99%, depending on the soil variables tested. RMSE values < 40% can be considered a good imputation in light of the low density of points used in this study. The study demonstrates strong capabilities for mapping forest soil properties at 250 m resolution, compared with the current Soil Landscape of Canada System, which is largely oriented towards the agricultural landbase. The methodology used here can potentially contribute to the national and international need for spatially explicit soil information in resource management science.

1. Introduction

Canada's forests span over 390 million ha of land, representing 10% of the world's forest cover and 30% of the world's boreal forest; mapping Canada's forest soil is therefore a huge challenge. Spatially explicit knowledge of Canada's soil properties is necessary not only for meeting national forest management needs but also for supporting participation in international environmental programs and studies.

In Canada, provinces and territories have a fiduciary responsibility for the stewardship of the natural resources within their administrative borders. Provincial forest inventory mapping programs follow standards and definitions specific to their respective jurisdictions, and soil information provided by these programs is difficult to harmonize across the country. Since the early 1900s, soil surveys have also been conducted for Canada's southern areas by a variety of federal and provincial agencies at scales ranging from 1:20,000 to 1:250,000 (Geng et al., 2010). Soil surveys traditionally focused on potential agricultural land for promoting the development of the most populated areas. Since the early 1980s, national soil maps and accompanying databases have been provided by the Soil Landscapes of Canada (SLC) as a product of the Canadian Soil Information Service (CanSIS), a component of Agriculture and Agri-Food Canada (Schut et al., 2011). The SLC database consists of 12,728 multi-component map units, with multiple taxonomic soil classes generalized from detailed soil surveys (Geng et al., 2010).

Although the latest public release of SLC version 3.2 (SLC, 2010) can be used for a variety of broad-scale spatial modeling applications, its main focus remains agricultural land with a relatively coarse resolution (1:1,000,000) using a vector mapping system (polygons) (Schut et al., 2011); no uncertainty or error assessment is provided. The greatest coverage of the Canadian commercial forest landbase can be found in earlier versions of SLC and has not been updated. The later SLC version provides coverage for areas with a mix of forest and agricultural lands (e.g. Atlantic Provinces). However, soil polygons in the forest landbase are very large and imprecise, and are therefore of limited use for forest policy and management decision-making.

There is a worldwide need for qualitative and spatial soil information for environmental monitoring and resource management (Hartemink et al., 2008; Lagacherie and McBratney, 2006; Sanchez et al. 2009). Soil is largely regarded as the foundation underlying a forest's capacity to provide environmental services; understanding, quantifying and monitoring soil patterns in relation to ecosystem health are therefore essential (Grunwald, 2009). Moreover, recent studies have shown how forest soil information can be used as predictive indicators of site

* Corresponding author.
E-mail address: Nicolas.Mansuy@RNCan-NRCan.gc.ca (N. Mansuy).

http://dx.doi.org/10.1016/j.geoderma.2014.06.032
0016-7061/Crown Copyright © 2014 Published by Elsevier B.V. All rights reserved.

suitability/sensitivity to intensive forest management practices such as biomass harvesting (Hazlett et al., 2014; Thiffault et al., 2014). Such functional linkage to forest nutrition and management increases the need for reliable and precise maps of soil properties in commercial forests. However, sparse or missing data is a common obstacle in soil mapping, and may lead to misrepresentations of soil characteristics and ultimately to inadequate or inappropriate resource management actions. In response to this situation, scientists around the world are increasingly turning to digital soil mapping and modeling approaches (DSMM) instead of conventional soil surveys (Grunwald, 2009; McBratney et al., 2003). DSMM is a broad class of soil map production methods that capitalizes on the availability of ancillary environmental predictive variables to generate maps of soil variables using a limited number of soil data. These methods are very explicit relative to conventional soil mapping. The predictive variables and their respective contributions are identified, error assessments are produced and results are repeatable.

The goal of this study was therefore to produce standardized grid maps of forest soils at a scale relevant for strategic-level reporting and decision-making at regional to national scales featuring selected soil variables (C, N, and soil texture) for upland forests across the Canadian commercial forest landbase using the DSMM approach. The overall aim was to define and test a methodology for soil mapping at 250 m resolution within the Canadian managed forest, which could then be adapted to meet other needs. The specific objectives were to: 1) identify the best soil covariates (i.e., environmental predictive variables) for spatially predicting soil variables in the forest floor and the top mineral horizons (to a 15 cm depth), 2) run the *k*-nearest neighbor method (*k*NN) for interpolating point source forest soil data across landscapes, and 3) evaluate the output of predicted (imputed) soil maps by performing internal validation using a leave-one-out method, and by comparing them with independent soil databases for the area of interest.

2. Materials and methods

2.1. Mapping area

The study area consists of 290 million ha of managed forests within seven ecozones of Canada illustrated in Fig. 1: Boreal Shield (BS), Atlantic Maritime (AM), Pacific Maritime (PM), Montane Cordillera (MC), Boreal Plains (BP), Boreal Cordillera (BC), and Taiga Plains (TP). The ecozones are at the top level of the Ecological Framework of Canada which defines the ecological mosaic of the country on a subcontinental scale (Ecological Stratification Working Group, 1996). The analysis was limited to the forested Canadian ecozones as defined by the National Forest Inventory and that have available and exploitable georeferenced soil legacy data. Therefore, Northern Taiga, the Arctic, the Prairies, the Mixedwood Plains and the Hudson Plains were not mapped. Moreover, this analysis only focused on upland forests and therefore did not cover areas dominated by agricultural land (according to the SLC classification) and wetlands (Fig. 1).

2.2. Data and analysis

In this study, a *k*NN algorithm was used to populate each cell (pixel) of a raster map corresponding to the study area with soil information, using georeferenced point source forest soil data. The *k*NN method is a non-parametric, multivariate approach to imputing observations or combinations of observations from sampling units, i.e., the reference set, to estimation or mapping units, i.e. the target set using

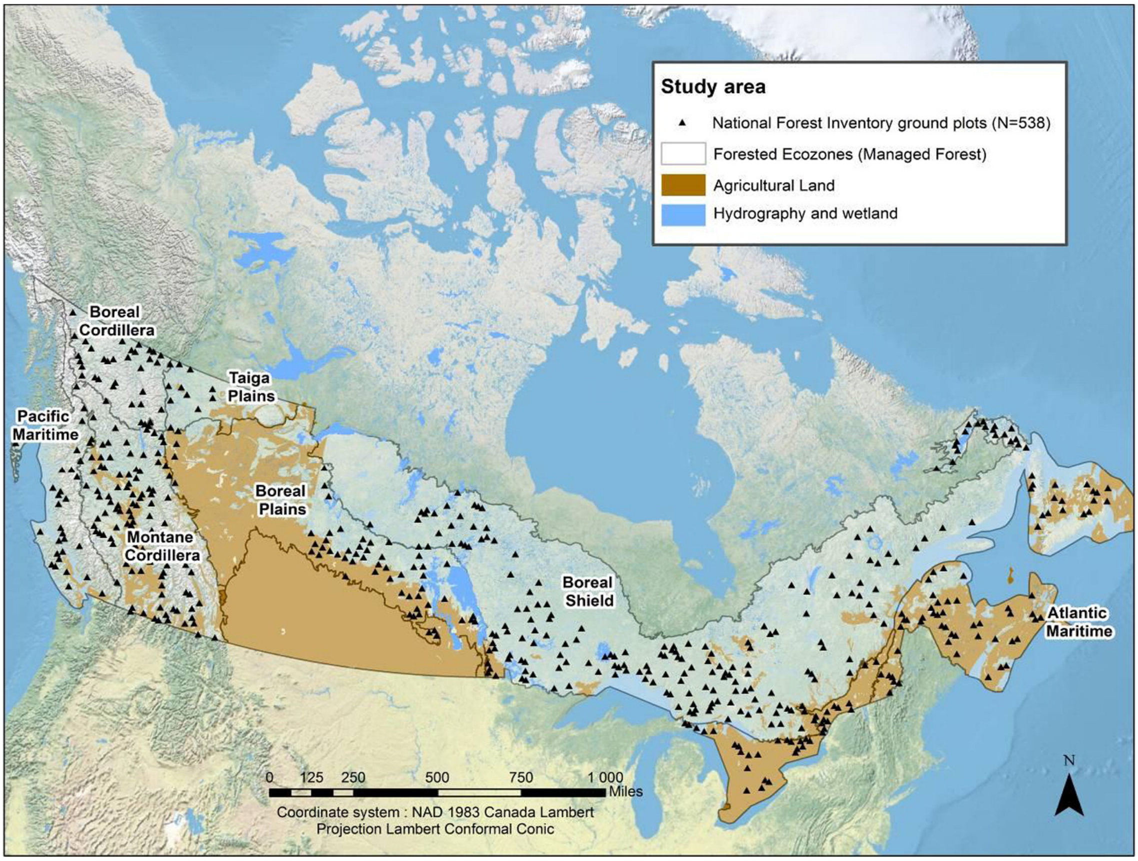


Fig. 1. National Forest Inventory (NFI) ground-plot data (538 points) across the Canadian forested ecozones. Areas dominated by agricultural land and wetlands were masked.

environmental variables (McRoberts et al., 2007). We used topographic and climatic information within the *k*NN analysis for interpolating point source forest soil data across landscapes. We also selected to use a 250 m × 250 m pixel corresponding to the grid of orthorectified mosaics from TERRA MODIS satellite imagery (Pouliot et al., 2009). This spatial resolution was selected because it is adequate to support strategic-level reporting at regional to national scales, and also because it ensured the direct compatibility of the digital soil products with similarly scaled Canadian forest inventory and forest change products (Beaudoin et al., 2014; Guindon et al., submitted for publication).

2.2.1. Reference soil variables

As a reference we used a set of physical and chemical soil variable measurements from georeferenced ground-plots (Fig. 1) collected as part of Canada's National Forest Inventory and made available for research use (NFI, Gillis et al., 2005). The use of the NFI ground-plot soil data as a reference dataset provided advantages for the *k*NN analysis in terms of quality/credibility of the information and of even spatial distribution of the sampling. All NFI plots are set up according to a systematic grid covering the Canadian forest landbase and sampled using a standard methodology, and soil and vegetation samples analyzed in the lab according to a standard protocol (Gillis et al., 2005). To date, approximately 1000 ground plots have been established.

For our study, only plots located on upland forest sites were selected among the NFI database. In NFI field plots, volumetric soil cores of the forest floor and the mineral layers at 0–15 cm, 15–35 cm and 35–55 cm depth, where possible, are collected. Since several NFI plots lacked samples from deeper soil layers (for example, due to the presence of bedrock), we restricted our study to the forest floor and the 0–15 cm mineral soil depth in order to have the most complete database possible for the greatest number of points. In total, 538 NFI ground-plots were included in the analyses. Four attributes of the forest floor (thickness; total nitrogen concentration; organic carbon concentration; carbon–nitrogen ratio) and six attributes of the upper 15 cm of the mineral horizons (proportion of sand, silt, and clay; bulk density; total nitrogen concentration; organic carbon concentration) were investigated in the analyses (Table 1). Forest floor total nitrogen and organic carbon concentrations were measured on oven-dried (at 70 °C) samples that were sieved with an 8 mm mesh and from which gravel and live roots were removed. Mineral soil texture, bulk density and total nitrogen and organic carbon concentrations were measured on the ≤2 mm air-dried portion of the samples. The choice of these soil attributes was justified by the fact that they have recently been identified as important indicators of site suitability for intensive forest management (Hazlett et al., 2014; Thiffault et al., 2010, 2014). We also restricted the analysis to the forest floor and the top layer of the mineral soil because they appear to be relevant for predicting a site response to forest management (Thiffault et al., 2010) and for capturing a significant part of the nutritional relationships between soil and trees in the studied ecosystems, since a large portion of the roots is located within the upper layers of soil (Jackson et al., 1996). The range of values for each soil attribute measured among the 538 NFI ground-plots is summarized by ecozone in Table 2.

2.2.2. Environmental predictor variables

In order to predict the soil attributes in target pixels, we used several raster datasets available at 250 m resolution that are likely to be correlated with soil properties in Canada. As generally suggested in the DSMM literature (e.g., McBratney et al., 2003), we first conducted a compilation of environmental covariates from a digital elevation model and from spatial climate models (Table 3). Ten topographic variables were derived from the USGS/NASA SRTM data and post-processed to provide continuous topography surfaces. We used eight climatic layers from the spatial climate models of McKenney et al. (2011). All raster layers were projected into Canada Lambert Conformal Conic and re-sampled if necessary to a 250 m resolution (1 pixel = 6.25 ha) pixel grid covering Canada's managed forest landbase. The predictor variables form the multi-dimensional feature space within which all reference and target pixels are located. The reference dataset was created by compiling the values of the 18 environmental predictor variables with the values of the 10 soil variables for the pixels within which the individual reference NFI plots were located. The pixels without NFI plots constituted the target set.

2.2.3. Pre-processing the reference dataset

We used the hybrid-Least Absolute Shrinkage and Selection Operator (hybrid-LASSO; hereafter HL; Efron et al., 2004) method to select the smallest subset of environmental predictor variables while retaining maximal predictive capacity, as in Beaudoin et al. (2014). The HL procedure was used to rank predictor variables according to the strength of their relationship with individual soil variables. These rankings were then merged into a single set, and environmental variables with consistently low rankings across all soil variables were removed. In addition, redundant environmental variables were removed when correlation between a pair of predictor variables was above 0.9.

2.2.4. The k-nearest neighbor (kNN) and mapping operations

Values of soil variables within the *k*-nearest neighbor pixels of the reference set were averaged and used as imputations to the target pixels. Formally, the estimate, $\tilde{y}_i$, for the *i*th target pixel was calculated as in McRoberts (2012):

$$\tilde{y}_i = \left(\sum_{j=1}^{k} w_j^i\right)^{-1} \sum_{j=1}^{k} w_j^i y_j^i \qquad (1)$$

where $\{y_j^i; j = 1, \ldots, k\}$ is the set of observations for the *k* reference set of pixels nearest in feature space built from the environmental variables to the *i*th target set pixel, as calculated using a given distance metric. The weights w_j^i used for each of the *k* pixels in the averaging process are calculated as:

$$w_j^i = d_{ij}{}^{-t} \qquad (2)$$

where d_{ij} is the distance in the feature space of the environmental variables between the *i*th target pixel and the *j*th nearest reference pixel calculated using a given distance metric (dm) and power *t* is used to

Table 1
Reference soil variables tested in the *k*NN mapping. All attributes are contained within the National Forest Inventory (NFI) ground plot circa 2001.

Soil variables	Soil names (abbreviations)	Units
Forest floor	Thickness (FFthickness)	cm
	Organic carbon concentration (FFOC)	g/kg
	Total nitrogen concentration (FFTN)	g/kg
	Carbon–nitrogen ratio (FFC:N)	Unitless
Mineral horizons	Clay content (Mclay)	%
(≤2 mm fraction in the 0–15 cm depth)	Silt content (Msilt)	%
	Sand content (Msand)	%
	Organic carbon concentration (MOC)	g/kg
	Total nitrogen concentration (MTN)	g/kg
	Bulk density (MBD)	g/cm^3

Table 2
Mean, minimum value (min), maximum value (max) and coefficient of variance (CV) of the soil variables by ecozone a) in the forest floor and b) in the first 15 cm of the mineral horizons. BS: Boreal Shield; BC; Boreal Cordillera; MC: Mountain Cordillera; AM: Atlantic Maritime; PM: Pacific Maritime; BP: Boreal Plains; TP: Taiga Plains.

a)	Thickness (cm)				Organic carbon (g/kg)				Total nitrogen (g/kg)				Carbon:nitrogen			
	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV
BS	12.5	1.5	40.0	64.0	413.2	119.0	540.5	17.0	12.1	4.9	20.8	27.3	37.4	16.3	76.4	34.5
BC	9.1	1.8	24.8	65.9	432.3	219.3	547.7	20.0	12.2	6.8	18.4	23.8	37.1	20.7	58.9	22.6
MC	6.3	1.2	32.3	77.8	374.0	81.6	522.5	24.8	11.1	5.7	17.2	23.4	34.7	17.2	61.2	23.6
AM	7.0	2.1	23.5	61.4	379.7	114.6	521.8	22.8	14.2	4.9	22.7	23.9	27.8	16.2	47.7	28.1
PM	11.7	1.5	27.5	63.2	464.8	241.1	548.1	14.9	12.7	8.7	23.5	29.9	38.4	20.1	59.7	30.5
BP	9.1	1.5	42.5	72.5	345.5	19.6	489.2	32.5	12.6	0.7	27.0	45.2	32.3	15.6	143.1	60.4
TP	12.8	4.3	31.5	61.7	390.3	63.2	541.8	36.5	16.0	9.4	22.7	30.6	32.4	17.6	57.8	39.8
Average	9.6	0.2	42.5	76.0	390.9	19.6	548.1	23.6	12.2	0.7	27.0	31.1	34.9	15.6	143.1	37.0

b)	Clay (%)				Silt (%)				Sand (%)				Organic carbon (g/kg)				Total nitrogen (g/kg)				Bulk density (g/cm^3)			
	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV	Mean	Min	Max	CV
BS	17.3	0.7	84.3	123.1	29.5	7.0	70.5	44.4	53.1	1.4	90.3	44.6	35.7	2.8	473.2	118.5	1.6	0.1	18.4	118.8	1.0	0.0	2.0	30.0
BC	15.4	5.1	28.2	32.5	37.3	19.1	60.8	30.3	47.4	21.3	75.8	29.5	37.7	8.7	145.1	81.7	1.8	0.2	5.8	83.3	1.0	0.4	2.0	30.0
MC	15.6	4.3	55.1	55.1	37.5	14.0	56.5	24.5	46.9	4.7	78.0	30.5	36.2	3.3	314.7	101.9	1.5	0.2	6.5	86.7	0.9	0.3	2.6	33.3
AM	11.1	1.9	41.2	70.3	36.8	14.4	66.3	32.6	52.1	18.2	80.0	31.1	39.8	8.3	160.9	73.6	1.9	0.6	5.1	57.9	0.7	0.1	1.3	42.9
PM	10.4	5.4	26.0	55.8	24.9	10.3	44.0	34.9	64.7	41.0	84.3	19.0	77.4	12.4	278.4	79.7	2.8	0.6	7.8	85.7	0.6	0.3	1.1	33.3
BP	20.2	1.8	69.8	78.7	26.2	3.0	66.4	52.3	53.6	14.0	95.3	42.2	23.4	3.8	150.0	116.2	1.3	0.2	6.8	115.4	1.3	0.0	1.8	23.1
TP	42.3	23.8	68.3	38.5	38.1	19.1	48.3	23.1	19.6	7.7	34.6	51.0	49.1	5.0	169.9	101.4	3.6	1.0	10.3	88.9	1.0	0.0	1.5	40.0
Average	16.4	0.7	84.3	100.6	31.3	3.0	70.5	41.2	52.2	1.4	95.3	40.2	37.3	2.8	473.2	105.1	1.7	0.1	18.4	100.0	1.0	0.0	2.6	40.0

CV = coefficient of variation (%) is defined as the ratio of the standard deviation to the mean.

modulate the impact of distance in the weighting (usually $0 \leq t \leq 2$). Based on the results of Beaudoin et al. (2014) and Thiffault et al. (2013), we used the Euclidean distance. In both of these studies, other distance metrics such as Mahalanobis (MAH) and Most Similar Neighbor (MSN) were found to provide either no or only a slight improvement in terms of fit relative to the Euclidian distance, with the latter also offering advantages over the other metrics in terms of simplicity of interpretation and derivation of error assessment.

Similar to Beaudoin et al. (2014), application of the *k*NN to the target pixel database was performed with an in-house C++ routine based on Approximate Nearest Neighbor Searching (ANN library, Mount and Arya, 2010) algorithm. The subsequent map editing of results was performed using ArcGIS 10.0 (Esri Inc.).

2.3. Testing the performance of the kNN method

The accuracy of the predicted soil values relative to observed values was assessed by leave-one-out cross validation using statistics computed by the same in-house routine described previously. The first method is the relative root mean square error (RMSE%) with the associated R^2, which provides an estimate of the standard deviation of the error relative to the predicted value (Eq. (3)). The second statistic is the relative bias (Bias%; Eq. (4)).

$$\text{RMSE\%}_m = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}}{\bar{y}} \times 100 \tag{3}$$

$$\text{Bias\%}_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)}{\bar{y}} \times 100 \tag{4}$$

where for a set m containing n pixels indexed i, $\tilde{y}_i$ is the *k*NN predicted value whereas y_i and $\bar{y}$ are the observed values. A greater predictive ability of the model is associated with a lower RMSE% and Bias%.

A third diagnostic statistic was used to assess the pixel-level accuracy and evaluate sampling representativeness and deficiencies using a variance estimator $\hat{\sigma}_i$ in the absence of spatial auto-correlation among the *k* neighbors for each soil variable (McRoberts et al., 2007). This metric was computed for each pixel *i* and expressed as a coefficient of variance (CV, %):

$$\text{CV}_i = \frac{\sqrt{\hat{\sigma}_i^2}}{\tilde{y}_i} \times 100 \tag{5a}$$

where

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{k}\left(y_{i,j} - \hat{\mu}_i\right)^2}{k-1} \tag{5b}$$

and where $\tilde{y}_i$ is the *k*NN imputed value, $y_{i,j}$ is the value of soil variable *i* for the *j*th neighbor and $\hat{\mu}_i = \tilde{y}_i$ is the estimated mean of the *k* neighbors. This measurement provides a spatially-explicit error assessment by highlighting areas where the dissimilarity between predicted and observed values is high and therefore where uncertainty in predicted values is highest.

In order to assess the performance of the *k*NN mapping method with independent datasets (not used in the analysis), the predicted proportions of clay, silt, and sand in the upper layers of the mineral soil were compared to values from two well-documented sources of information on Canadian soil properties, i.e. the Soil Landscapes of Canada (SLC), and maps of surficial deposits prepared by the ministère des Ressources naturelles du Québec (MRN, 2013). These two sources were considered

Table 3
Names and types of the raster layers used as predictor variables. All the layers have 250 m resolution in the grid (raster) format. The climatic layers correspond to the 1970–2000 time period (McKenney et al., 2011). The topographic layers are derived from the USGS/NASA SRTM data computed using ArcGIS 10.0.

Types	Names or abbreviations	Units	Definitions
Climatic	ACMI	cm/year	Annual moisture index. annual ACMI = P − PET. P is the annual precipitation. PET is the annual potential evapotranspiration (loss of water vapor from a well-vegetated landscape).
	SCMI	cm/summer	Summer moisture index. summer SCMI = P − PET. P is the summer precipitation. PET is the summer potential evapotranspiration (loss of water vapor from a well-vegetated landscape).
	PWQ	mm	Precipitation of the warmest quarter. The warmest quarter of the year is determined (to the nearest month) and the total precipitation over this period is calculated.
	THM	°C	Mean maximum daily temperature of the hottest month. The highest temperature of any monthly maximum temperature.
	TCM	°C	Mean minimum daily temperature of the hottest month. The lowest temperature of any monthly minimum temperature.
	MAT	mm	Mean annual precipitation
	TAP	mm	Total annual precipitation. The sum of all the monthly precipitation estimates.
	PCQ	mm	Precipitation of the coldest quarter. The coldest quarter of the year is determined (to the nearest month) and the total precipitation over this period is calculated.
Topographic	Elevation	Meter	Digital elevation models obtained from the Shuttle Radar Topography Mission (SRTM)
	Aspect	0 to 360°	The downslope direction of the maximum rate of change in value from each cell to its neighbors. Aspect can be thought of as the slope direction. Aspect is expressed in positive degrees from 0 to 360, measured clockwise from north (flat = −1)
	Beers aspect	0 to 2	Heat index for use in predicting forest productivity; Beers aspect = 1 + cos((45° − aspect) / slope_deg)
	Slope	%	The rate of maximum change in z-value from each cell.
	Profile curvature	−0.189 to 0.417	The profile curvature is in the direction of the maximum slope, affects the acceleration and deceleration of flow and so influences erosion and deposition.
	Relative moisture index	0 to 6.8	Relative moisture index (RMI also referred to as wetness index) = relative amount of water flowing into a pixel (flow accumulation) in relation to amount flowing out based on slope. RMI is very similar to the topographic convergence index.
	Watershed stream	1 to 1831	The local drainage area enclosed between the local divide and the stream into which each cell drains.
	Flow direction	1 to 128	The flow direction from each cell to its steepest downslope neighbor.

truly independent from the data used in our study since they both predate the collection of the NFI soil samples, and were developed by different organizations. Proportions of clay, silt, and sand were chosen as examples in this exercise because they are the most common soil attributes documented for Canada as a whole.

As the first source of comparison, the SLC database provides soil textural classes at the level of the ecodistrict (SLC, v2.2, 1996). Ecodistricts are the smallest units (minimum size ~ 100,000 ha) of the National Ecological Framework (Ecological Stratification Working Group, 1996); they are mapped at a scale of 1:2 M and are characterized by relatively homogeneous biophysical and climatic conditions. The comparison involved averaging the *k*NN predicted values of percent clay, silt and sand of the pixels within each SLC ecodistrict polygon for which soil textural class was available in the SLC database. In total, 215 out of 478 ecodistricts were used (~100 million ha or 35% of the area of interest) of which 71 were dominated by sand, 48 by clay and 96 by loam.

As the second source of comparison, the map of surficial deposits in Quebec was used. This vector map provides knowledge on the origin, mode of formation and material matrix of surficial deposits, which may act as the proxy for the expected granulometry of the soil at a 1:40,000 scale, which is close to the scale of the *k*NN maps. The comparison involved averaging the *k*NN predicted values of percent clay, silt and sand of the pixels within each surficial deposit polygon for the portion of the Quebec province area that fell within our study area.

3. Results

3.1. Selection of the best predictor variables and parameter optimization

Based on the HL analysis, twelve of the original 18 predictor variables tested were retained (Fig. 2a): 7 topographic (elevation, slope, Beers aspect, profile curvature, relative moisture index (RMI), watershed stream and flow direction) and 5 climatic variables (annual moisture index (ACMI), precipitation of the warmest quarter (PWQ), mean maximum daily temperature during the hottest month (THM), mean minimum daily temperature during the hottest month, mean annual precipitation (MAT)). These variables consistently showed the strongest relationships with individual soil variables and were not redundant among them.

Using this smaller set of predictor variables, the HL revealed that mineral soil bulk density (MBD) was the predicted variable showing the highest proportion of explained variance with an adj. $R^2 > 0.20$, followed by proportion of silt in mineral soil (Msilt; adj. $R^2 > 0.15$), thickness of forest floor (adj. $R^2 > 0.10$), and then proportions of clay (Mclay) and sand (Msand) in the mineral soil and organic carbon concentration in the mineral soil (MOC) (adj. $R^2 > 0.05$). The THM and the elevation explained the most variance for forest floor thickness. The proportion of the variance explained by the two types of environmental predictor variables (climatic or topographic) differed widely between each soil variable (Fig. 2b). For example, climatic variables accounted for 100% of the explained variance for forest floor organic carbon concentration (FFOC), while topographic variables accounted for almost 70% of the variance for total nitrogen concentration in the mineral soil (MTN). On average, for all soil variables, topographic predictors accounted for 37% of the explained variance versus 64% for the climatic predictors.

In order to select the optimal number of *k* nearest-neighbors for imputing soil attributes, multiple *k*NN analyses were run with the 12 best predictors using the Euclidean distance metric. We found $k = 10$ to be optimal for the entire soil dataset as shown by the leveling of the RMSE% (Fig. 3).

3.2. kNN performance and cartography of the selected soil variables

Table 4 shows the summary statistics of the predicted and observed values as well as the RMSE% values for all the variables. The RMSE% values for the predicted soil variables ranged from 22.37 to 98.92 (Table 4). The soil variables with a value of RMSE% < 40% included forest floor organic carbon (FFOC), total nitrogen concentrations in the forest floor (FFTN), MBD and sand content (MSand), forest floor C:N (FFC:N) and mineral soil silt content (MSilt). Values of RMSE% for the other soil variables such as the mineral soil clay content (MClay) were between 74.73 and 98.92. Scatter plots for the ten soil variables showed an overestimation at low observed values and an underestimation at the high ones (Fig. 4).

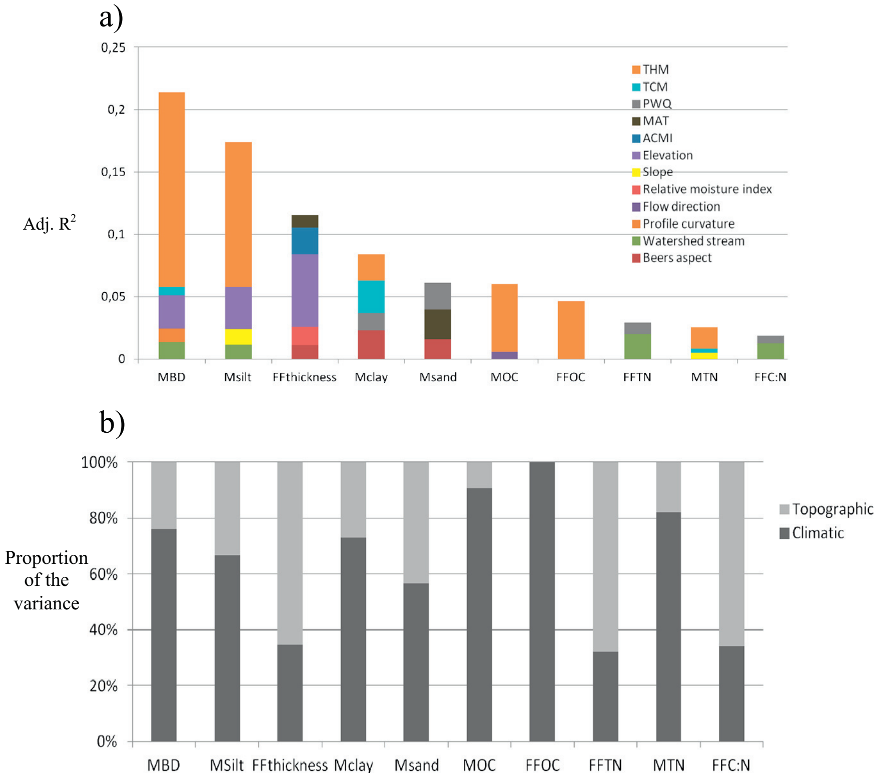


Fig. 2. a) Values of adjusted R^2 from the hybrid-LASSO (HL) selected predictor variables (N = 538 pts) for each soil variable. See Table 1 for the definitions of the soil variables and Table 3 for the predictor variables. b) Proportion of each type of predictor variable explaining the adjusted R^2 from the HL for each soil variable tested.

Example maps for soil attributes with value of RMSE < 40%, and their associated maps of coefficient of variance, are presented and discussed hereafter (the remaining maps are available in Supplementary data). Forest floor organic carbon concentration showed a relatively homogeneous pattern across the ecozones (Fig. 5a), with predicted values between 106 and 500 g/kg and an average of 389 g/kg. However, the northwestern part of the Boreal Shield and to some extent the southern portion of the Taiga Plains ecozones showed lower values between 200 and 250 g/kg. However, this zone was also associated with high rates of error, as expressed by the coefficient of variance (Fig. 5b) likely as a result of a particularly low density of reference plots.

The proportion of sand in the mineral soil showed predicted values ranging between 5 and 94% and an average of 52% (Fig. 5c). The Pacific Maritime, Boreal Shield and Montana Cordillera ecozones showed average values of approximately 40%, while the area in the center of the Boreal Shield ecozone had lower estimated values between 30 and 40%. However, low values within the Boreal Shield ecozone were also associated with a high rate of error (Fig. 5d). Nevertheless, the map makes it is possible to detect the Clay Belt formation, an area with very low concentration of sand in the northernmost part of the Boreal Shield. The Clay Belt is a physiographic unit composed mostly of glaciolacustrine deposits left by the proglacial Lake Ojibway, stretching between the Cochrane District in Ontario, and Abitibi County in Quebec, covering 120,000 km^2 (Veillette, 1994).

The C:N in the forest floor showed predicted values ranging between 20 and 70 with an average of 35 (Fig. 5e). The Montane Cordillera,

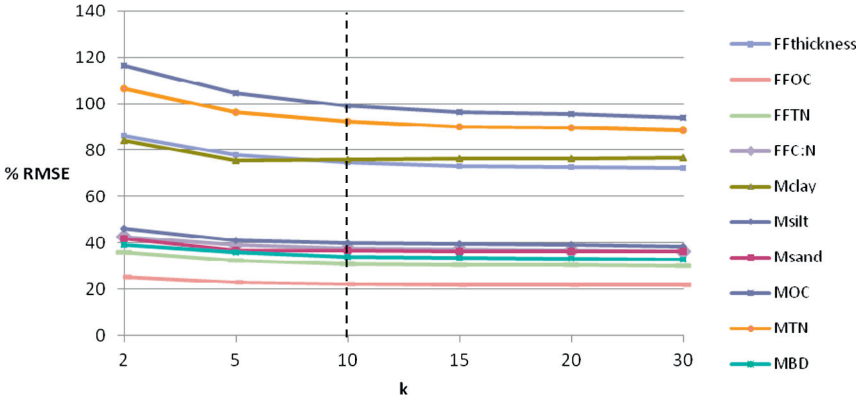


Fig. 3. Variation of the RMSE (%) for various *k* values based on the entire dataset (N = 538). *k* = 10 was retained to run the *k*NN mapping procedure. Definitions of soil variables are explained in Table 1.

Table 4
Assessment of results for all ecozones together, ranked in increasing order according to the RMSE value (%). Abbreviation of the soil variables are in Table 1. The statistics are described in the text (Eqs. (3) and (4)).

Variables	Predicted minimum value	Predicted maximum value	Predicted mean value	Predicted standard deviation value	Relative Bias (%)	RMSE (%)	R^2
FFOC (g/kg)	106.83	501.9	389.05	53.28	0.53	22.37	0.143
MBD (g/cm^3)	0.36	1.66	0.96	0.22	−0.00	34.14	0.175
FFTN (g/kg)	0.68	20.79	12.15	2.94	0.08	35.82	0.025
Msand (%)	5.97	94.25	52.28	12.55	0.14	36.68	0.195
C:N	21.42	70.73	34.77	6.46	−0.01	37.82	0.056
Msilt (%)	12.97	57.5	31.27	7.68	−0.32	39.88	0.131
FFthickness (cm)	2.00	26.08	9.74	3.99	−0.22	74.73	0.085
Mclay (%)	2.12	68.05	16.44	11.6	0.15	76.07	0.431
MTN (g/kg)	0.20	5.22	1.67	0.604	1.54	88.38	0.052
MOC (g/kg)	4.99	144.08	36.37	17.49	0.59	98.92	0.199

Boreal Cordillera and eastern Boreal Shield ecozones had values ranging from 25 to 30, while the northwestern part of the Boreal Shield and the Taiga Plains ecozones had values above average between 40 and 50. The C:N error map was similar to that of the organic carbon in the forest floor, with highest levels of error in the northern Boreal Plains and Taiga Plains ecozones (Fig. 5f).

3.3. Mapping validation with two independent datasets

Maps of soil attributes were first compared with textural classes attributed to ecodistricts of SLC. The predicted proportions of clay, silt and sand showed coherence with the SLC ecodistrict-level textural classes (Fig. 6). The average predicted proportion of clay obtained with the *k*NN procedure was higher in ecodistricts with a clay textural class (23%) compared with ecodistricts with a sandy or loamy textural class (Fig. 6a). The *k*NN predicted proportion of sand was higher (63%) in the ecodistricts with a sand textural class compared with ecodistricts with the clay (43%) or loam (42%) textural class (Fig. 6c). However, it could be noted, by looking at the red crosses, which represent observed soil textural values from NFI plots, that the predictions overestimated sand and silt contents in areas with SLC-determined low silt content and underestimated clay content in areas with SLC-determined high clay content (Fig. 6).

When compared with the map of surficial deposits for the province of Quebec, the *k*NN predicted proportion of clay increased from the coarser deposits mainly composed of sand (e.g. littoral marine and fluvioglacial) towards the fine particle deposits such as glaciolacustrine deposits (Fig. 7a; Appendix B). The predicted proportion of sand also decreased from the coarser deposits towards the finer particle deposits (Fig. 7c). The comparison with surficial deposits therefore suggested a good coherence between the two sources of data.

4. Discussion

4.1. Improvements of soil mapping in Canada

This study represents the first national digital soil mapping for the managed forest in Canada. Despite the low density of points available in the reference set, with only 538 points for all of Canada, the results are a marked improvement in terms of resolution and coverage of the forest landbase in comparison with the SLC which is largely focused on agricultural land and therefore provides only crude information about the forest landbase. The SLC is also limited by its coarse resolution across the forested landbase and by its vector mapping system (Schut et al., 2011). By providing a methodology for mapping the forest soil, our study is complementary to existing efforts expanded towards the production of digital soil products for Canada by SLC proponents and by others involved in soil mapping.

With pixels of 6.25 ha compared with polygons of 1000 ha in SLC, our results also show the capacity of the methodology used to produce standardized moderate resolution soil digital maps across land cover types and jurisdictional boundaries (provinces and territories), even where the point source data are sparse. Moreover, maps in raster format offer a range of advantages over conventional, vector-based soil maps, as identified by Hartemink et al. (2008). These include flexibility for quantitative studies such as C balances, the shift from soil texture classes to continuous variables, the ease of integration with other raster-based data, and the provision of uncertainty and error assessments.

In our study, we chose to use a small but well-distributed set of field-based soil point data to populate the reference dataset and impute values across landscapes, which also allowed for proper quality assessment/quality control of soil property values. Many studies on soil mapping use data extracted from existing vector soil maps, which are then used as reference to predict soil variables and sometimes to validate the outputs as well (e.g., Bui and Moran, 2001, 2003; Heung et al., 2014). Vector soil maps have the advantage of being easier to access than point data, and provide useful information about soil–landscape relationships. However, map units are generally not pure, i.e. the whole area of the polygon is likely not homogenous. This is an important issue for the large SCL polygons in the forest landbase in which disaggregation into its individual map components is problematic because of the lack of within-polygon spatial information (Bui and Moran, 2001; Odgers et al., 2014). Digital soil mapping based on legacy soil maps therefore carries both the caveat of relying on the mental models developed by the pedologists who produced the maps, and the uncertainty associated with the disaggregation process. On the other hand, field point data are more objective in nature than legacy soil maps. However, they may not be totally representative of the pixel size chosen for the digital soil mapping. Nevertheless, legacy soil maps for the Canadian forest landbase, or other form of land classification, could be used as predictors or as stratification in further exercises to see if they improve the prediction fit.

4.2. Performance and limits of the method

The *k*NN method is becoming increasingly popular for mapping forest attributes (Bernier et al., 2010; Franco-Lopez et al., 2001; McRoberts, 2012; Tomppo et al., 2008; Wilson et al., 2012). It has also been applied successfully in soil sciences to estimate soil particle-size distribution and water regimes (Nemes et al., 1999, 2006), and the availability of soil base cations (Thiffault et al., 2013). Its non-parametric nature makes it well suited to capture complex and non-linear relationships, because estimates are made from local subsets of data rather than relying on globally optimized relationships (Nemes et al., 2006). Moreover, maps can be rapidly and efficiently updated as new field measurements become available. Other machine-learning methods have also been shown useful for soil DSM, such as Random Forest (e.g. Heung et al., 2014), Multiple Additive and Regression Tree (e.g. Lacoste et al., 2011) and artificial neural networks (e.g. Behrens et al., 2005); they are similar in nature to *k*NN as they are all data-mining methods for knowledge discovery based on algorithms aimed at learning relationships between response variables and predictors (Breiman, 2001).

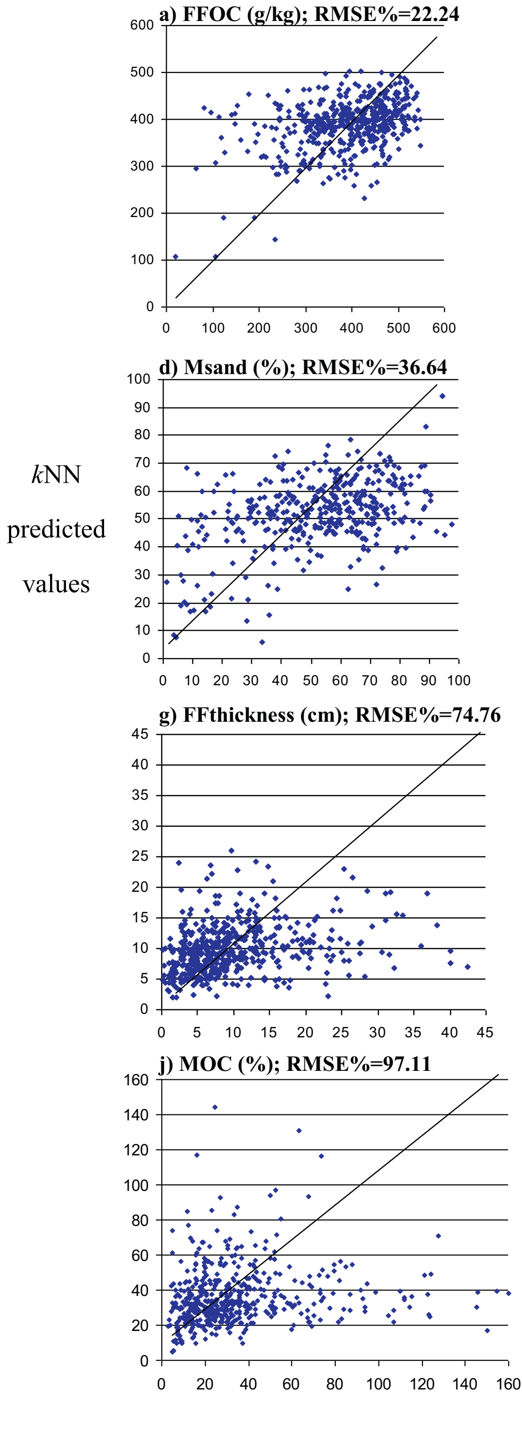
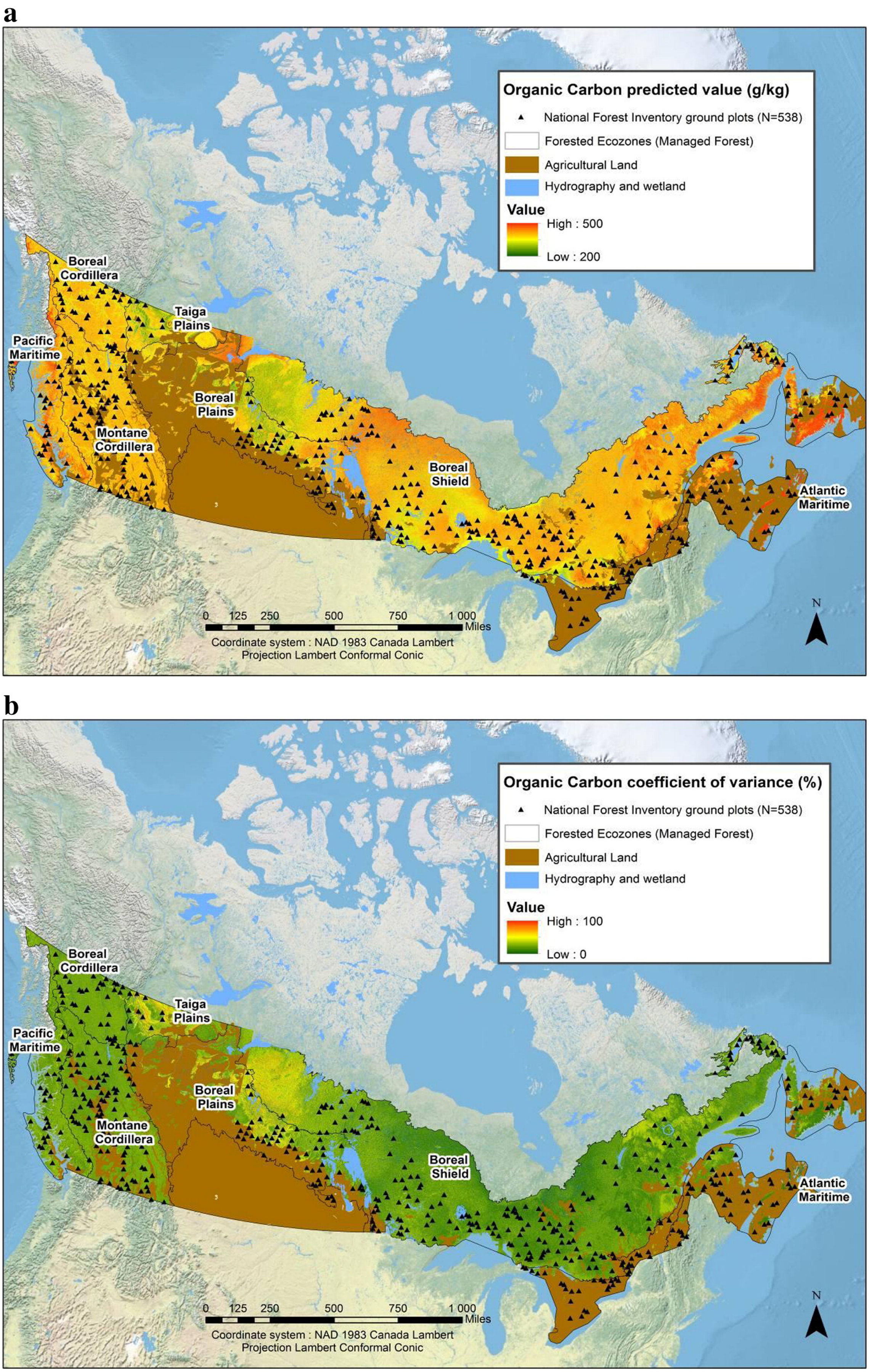


Fig. 4. Observed versus predicted values for all ecozones together for the soil variables: a) concentration of organic carbon in the forest floor; b) concentration of total nitrogen in the forest floor; c) bulk density in the mineral soil; d) proportion of sand in the first 15 cm of mineral horizons; e) C:N in the forest floor; f) proportion of silt in the first 15 cm of mineral horizons; g) thickness of the forest floor; h) proportion of clay in the first 15 cm of mineral horizons; i) concentration of total nitrogen in the first 15 cm of mineral horizons and j) concentration of organic carbon in the first 15 cm of mineral horizons. The 1:1 diagonal is the line of perfect prediction. The units are not all consistent.

Given the low density of reference point data in the training dataset (i.e. 538 points), predicted maps with RMSE values < 40% may be considered operational and sufficient to provide useful knowledge for the whole of the Canadian managed forest landbase. Moreover, the external validation with an independent dataset at the province of Quebec scale shows good coherence between the predicted values of the textural classes and the different types of surficial deposits, given the 48 training points available in the province. The only study comparable in scope

Fig. 5. Maps depicting three predicted soil variables within the managed forest and their associated maps of coefficient of variance (Eqs. (5a) and (5b) in the text). a) and b) concentration of organic carbon in the forest floor; c and d) proportion of sand in the first 15 cm of mineral horizons; e and f) C:N in the forest floor. For better display of contrast, the ranks of value (min–max) displayed may be different from those in Table 4.


a
Organic Carbon predicted value (g/kg)
▲ National Forest Inventory ground plots (N=538)
□ Forested Ecozones (Managed Forest)
🟫 Agricultural Land
🟦 Hydrography and wetland
Value
High : 500
Low : 200
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000
Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N
b
Organic Carbon coefficient of variance (%)
▲ National Forest Inventory ground plots (N=538)
□ Forested Ecozones (Managed Forest)
🟫 Agricultural Land
🟦 Hydrography and wetland
Value
High : 100
Low : 0
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000
Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N

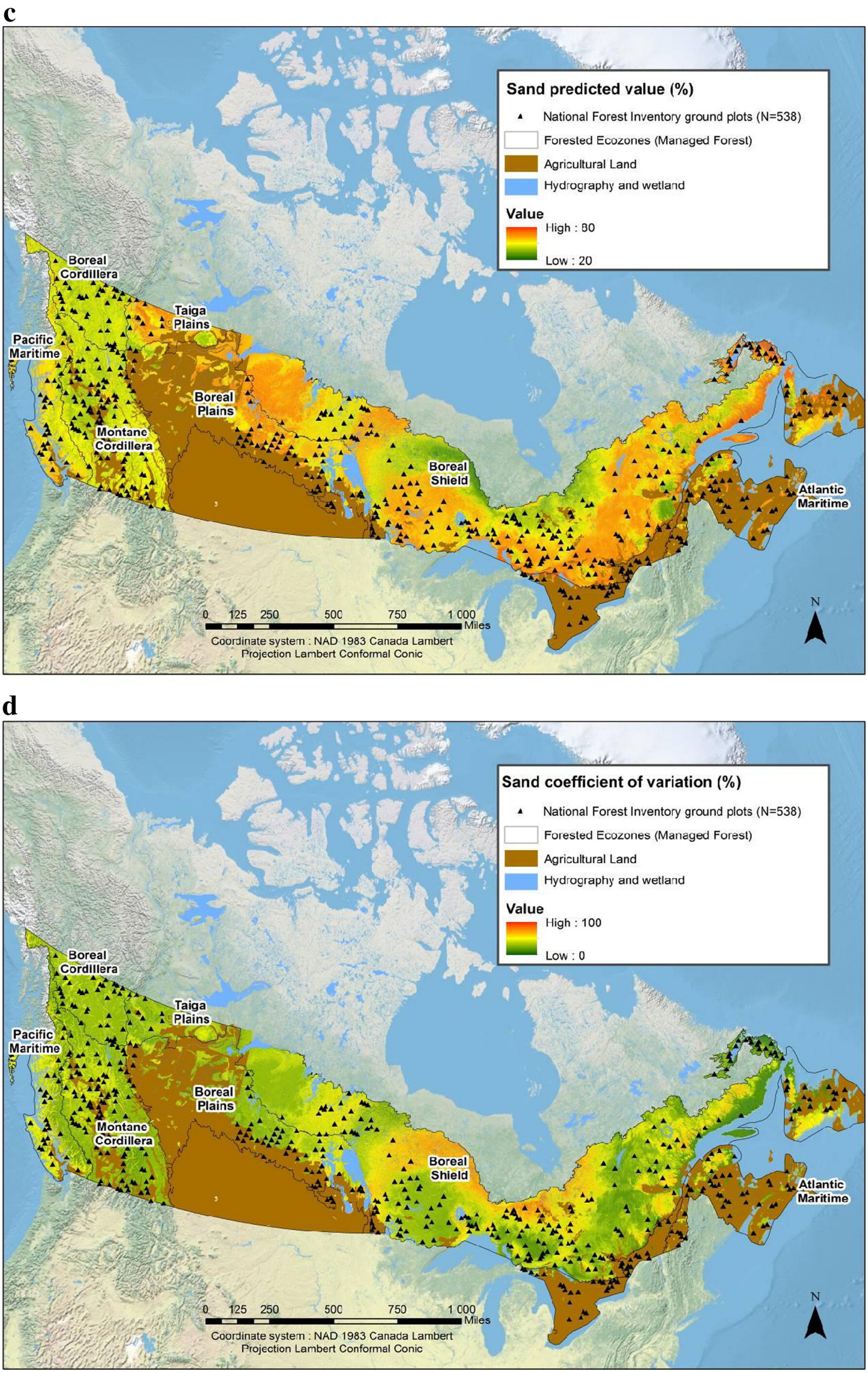

c
Sand predicted value (%)
National Forest Inventory ground plots (N=538)
Forested Ecozones (Managed Forest)
Agricultural Land
Hydrography and wetland
Value
High : 80
Low : 20
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000 Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N
d
Sand coefficient of variation (%)
National Forest Inventory ground plots (N=538)
Forested Ecozones (Managed Forest)
Agricultural Land
Hydrography and wetland
Value
High : 100
Low : 0
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000 Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N

Fig. 5 (*continued*).

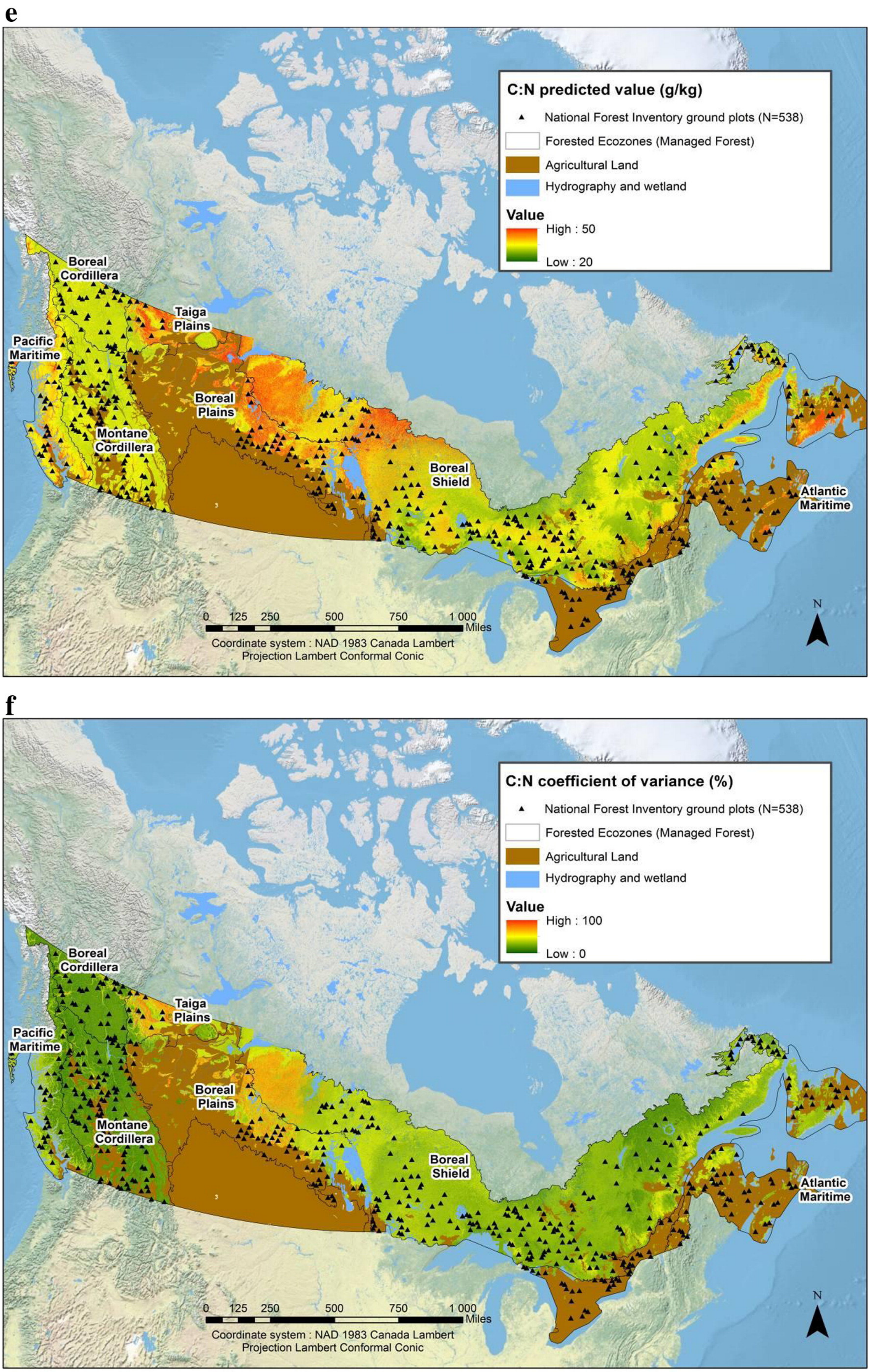

e
C:N predicted value (g/kg)
National Forest Inventory ground plots (N=538)
Forested Ecozones (Managed Forest)
Agricultural Land
Hydrography and wetland
Value
High : 50
Low : 20
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000 Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N
f
C:N coefficient of variance (%)
National Forest Inventory ground plots (N=538)
Forested Ecozones (Managed Forest)
Agricultural Land
Hydrography and wetland
Value
High : 100
Low : 0
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
Atlantic Maritime
0 125 250 500 750 1 000 Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic
N

Fig. 5 (*continued*).

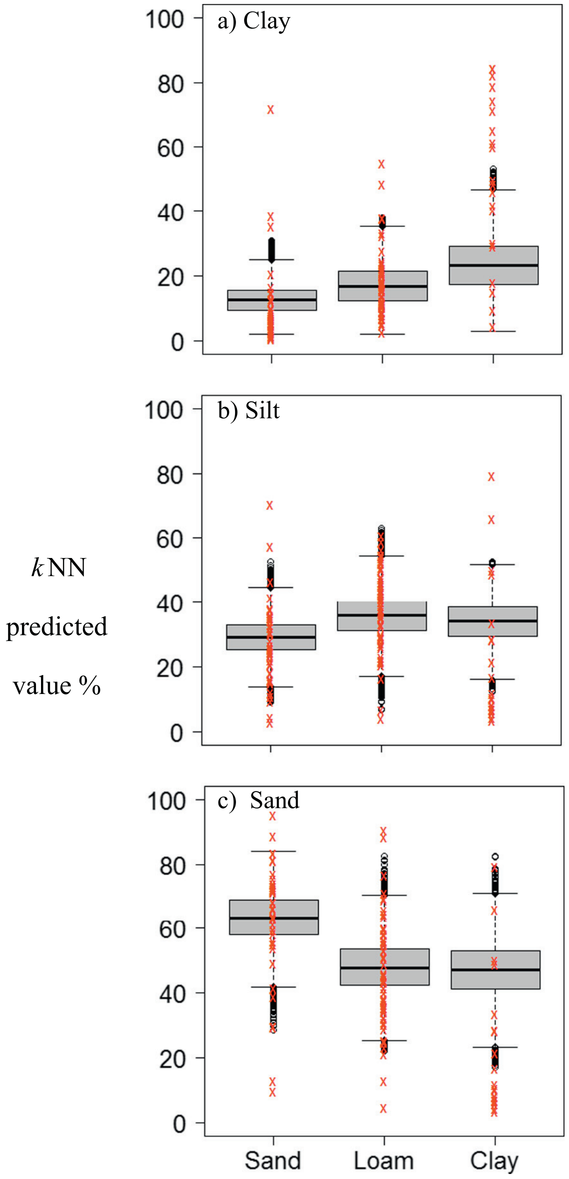


Fig. 6. Boxplots of the predicted values of proportions of particle sizes in the first 15 cm of mineral horizons for a) clay b) silt, and c) sand within the sand, loam and clay textural classes in the ecodistrict across the managed forest of Canada. The red crosses show the observed values in each textural class. The soil textural classes were gathered from the Soil Landscape of Canada.

and methodology to ours (Henderson et al., 2005) found RMSE% values of predicted proportions of clay and organic carbon at 250 m resolution in Australia in the same order as ours.

Accuracy of prediction for maps with a RMSE > 40% may be questionable and precautions are needed when using these estimated values. Accuracy of prediction is also limited for areas with low RMSE%, because imputation methods such as *k*NN generally tend to overestimate the lowest values and underestimate high values (Fig. 4), which decrease the range of variability for any distribution (Scheuber, 2010; Suominen et al., 2013). This overestimation/underestimation of values in the extremes of the range is common in *k*NN studies (Beaudoin et al., 2014; Magnussen et al., 2010), because this method is prone to bias when the values of the predictors for a unit in the target set do not have close neighbors or fall outside the range of values in the reference set. The prediction for this target unit will still be that of the closest neighbor in the reference set (however distant it might be), with the net effect of underestimating large values and overestimating small ones (Magnussen et al., 2010). Avenues to overcoming this bias include: increasing the representativeness in the reference set of the whole range of predictor values among the target set, using land stratification, and implementing bias correction methodologies if possible (Magnussen et al., 2010).

In fact, some precautions are necessary in DSMM because there is consistently an error between the values of variables observed from ground plot measurements and the corresponding predicted variables derived from digital maps (Xu et al., 2009). The first source of error can often emanate from the choice of predictor variables. Indeed, the selection of predictor variables with a low R^2 in this study (Fig. 2) may be statistically doubtful. Despite the low R^2, the environmental predictors selected are likely to be relevant for prediction of soil attributes as they are part of the empirical factors that are commonly related to the soil formation in reference to SCORPAN (Soil; Climate; Organisms; Relief; Parent material; Age; Space; Behrens et al., 2010; Boettinger, 2010; Minasny and McBratney, 2010) especially in the boreal forest (Seibert et al., 2007). The R^2 values from environmental predictors for each soil variable (i.e. Fig. 2) should be interpreted with caution since the multivariate non-linear nature of *k*NN does not provide direct causal relationships between two variables.

Our study showed good performance at predicting soil characteristics in the mineral layer, which is somewhat different than what is found in other studies, where prediction of lower soil layers is often poor (e.g. Henderson et al., 2005). Notably, our results provided good accuracy of prediction (low RMSE%) to MBD. This may be due to the fact that MBD integrates the variability of several soil properties including clay, silt, sand and organic matter content (Sequeira et al., 2014).

Nevertheless, results from our study displaying the links between environmental predictors and soil variables are aligned with findings from previous studies. For instance, the fact that organic carbon concentration in the forest floor and the upper mineral horizons which is almost entirely related to variations in temperature in our study is similar to conclusions with other large-scale studies that show that climatic factors, notably temperature, largely drive soil carbon concentration over large scales (e.g. Ladd et al., 2013; Post et al., 1982). Also, forest floor thickness has been intensively studied in boreal regions, notably due to its link to paludification, and it was shown that factors influencing water availability and drainage explain forest floor thickness (e.g. Laamrani et al., 2014; Seibert et al., 2007). This is similar to findings from our study, in which the amount of precipitation and relative moisture index were important predictors for this soil variable. Links between topographical parameters related to water movement and forest floor nitrogen such as those seen in our study (forest floor total nitrogen concentration and C:N are related to variations in watershed stream, i.e. local drainage area) are also found in the literature (Seibert et al., 2007; Welsch et al., 2001). Nevertheless, a further study should look at broadening and refining the choice of environmental predictors to increase the power and accuracy of digital soil mapping across Canadian landscapes. For instance, bedrock composition has been shown to be a powerful covariate for predicting soil properties (Gray et al., 2014) and for explaining variations in vegetation (Hahm et al., 2014).

The second source of error can be attributed to the low density of points in the reference dataset (i.e. 538 points). The overall low adj. R^2 is more than likely due to the low number of NFI points available for the analysis compared with the number of variables predicted. Uncertainties may also arise from complex interactions between environmental variables combined with limited representation of the variability (narrow range) of soil properties and related soil forming processes among the training data. A common problem of DSMM is the difficulty to adequately predict properties of rare soils or of areas with high complexity. This problem could be overcome by using training sets with better representation of unique/uncommon features. However, as stated by Heung et al. (2014), the selection of data for training sets should reflect the goal of the study, either to maximize overall prediction accuracy over landscapes, or to maximize prediction accuracy of rare

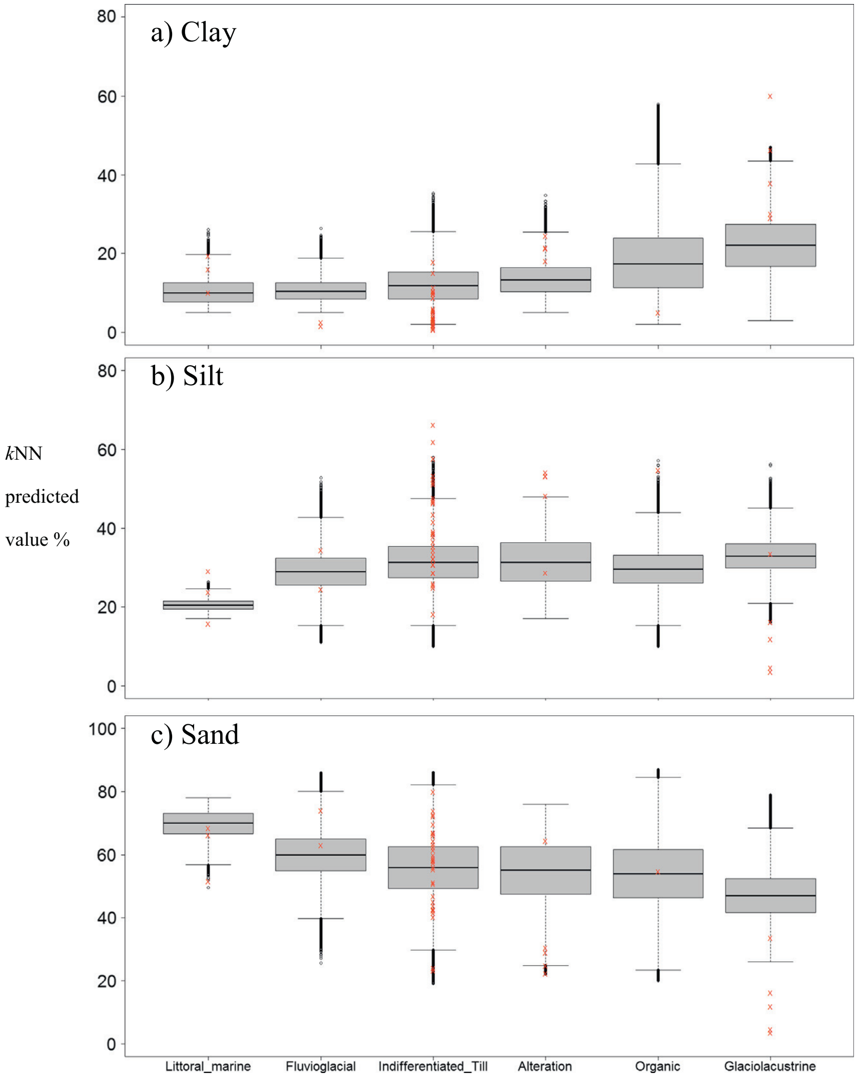


Fig. 7. Boxplots of the predicted values of proportions of particle sizes in the first 15 cm of mineral horizons for a) clay, b) silt and c) sand within the major surficial deposit types in the province of Quebec. The red crosses show the observed values in each type of surficial deposit (N total = 48). The surficial deposit types were gathered from the 1/40,000 provincial forest inventory map (modified from Mansuy et al. 2011) and are described in Appendix B.

occurrences. In fact, given that the quality of the results is directly related to the ancillary data and that the *k*NN method gives only results inherent to the reference dataset, improvement in the prediction of soil properties does not rely on more sophisticated quantitative methods, but rather on gathering more, more useful, higher quality and better-distributed data, as suggested by numerous DSMM studies (i.e., Grunwald, 2009; Grunwald et al., 2011). Numerous databases of soil points exist across Canada, and given that proper standardization and quality assessment/quality control can be performed on them, they will be useful for improving the predictive power of future mapping exercises.

4.3. National and international perspectives for better monitoring resources

The results provided in this study fill a gap in terms of providing soil maps for the Canadian forest landbase and focus on soil properties and soil depth that have the greatest relevance for forest nutrition and site response to forest management. Moreover, the methodology tested here can be used for the standardization of soil information at the national scale for a better monitoring of environmental resources in Canada. The method developed in this study demonstrates a shift from traditional soil mapping towards the new digital soil mapping approaches as suggested by Geng et al. (2010). The comprehensive set of maps could be useful in improving the delineation of the ecological stratification units. For example, the soil layers obtained in this study can be supplemented with layers of vegetation attributes also imputed with the *k*NN method and MODIS imagery at 250 m by Beaudoin et al. (2014) and offer standardized attributes of managed forests in Canada. Digital soil mapping could significantly reduce uncertainties by helping to address a wide range of national environmental challenges encountered in Canada such as natural disturbance management (fires and pests), monitoring the forest carbon budget, or the development of the biomass sector.

From an international perspective, this study contributes to satisfying the global need for accurate, up-to-date and spatially referenced soil information, as identified by the Global Digital Soil Properties Consortium (www.globalsoilmap.net). A global soil map network is being developed to provide soil information to meet the demand of a broad range of users including multiple levels of government, natural resource managers, educational and research institutions, and agriculturalists (Arrouays et al., 2014). The methodology of our study can be adapted to meet other requirements in terms of resolution, e.g.

producing maps at a finer resolution such as the 100 m pixel used in Global Soil Mapping, or extending to deeper soil layers and to other soil variables. Moreover, our methodology based only on soil pit data and easily accessible DEM derivatives and climate information makes it well suited for trans-frontier mapping. Thus, we strongly believe that the future development of the products generated and discussed in this study could complement the current global mapping efforts.

5. Conclusion

This study represents a successful initial step in the development of digital soil maps using *k*NN technique of imputation. The main goal was to improve and standardize soil data for the managed forests of Canada without the constraints of the various provincial and territorial jurisdictional boundaries. Because this method offers the possibility of mapping soil variables at a 250 m resolution grid, this implies considerable improvement over the previous soil programs which mapped only broad multicomponent soil type polygons. However, these results can be considered preliminary and there is room for improvement in providing up-to-date and spatially explicit soil information tailored to the end user's needs. For instance, the need for a national database of soil sampling sites is clearly identified. Another area of development would be to generate smaller pixels to minimize the scale mismatch between ground plots and targeted pixels (Xu et al., 2009). Therefore, future efforts will attempt to incorporate various soil attributes, organic and/or mineral, at various depths within soil profiles for pixels at 30 m resolution in order to satisfy the objectives of the GlobalsoilMap.net international project.

Acknowledgments

This project was made possible through funding from the Program for Energy Research and Development (TBC017-1314) and the EcoEnergy Innovation Initiative of Natural Resources Canada (BIOI027), as well as funding from the Canadian Boreal Forest Agreement. We would like to acknowledge the contribution of Dan McKenney and his team at the Canadian Forest Service, Natural Resources Canada for providing the spatially explicit rasterized climate-related feature variables. We express our sincere thanks to Canadian Forest Service programmer Rémi St-Amant for the creation of a fast and efficient C++ *k*NN mapping code, without which the task of performing Canada-wide analyses would have been a significantly greater challenge. We thank all partner agencies for their ongoing collaboration in the NFI program, which provided the data foundation for this research. We are also thankful to the ministère des resources naturelles du Québec (MRN) for the surficial deposit map. Darren Allen, of Resources Natural Canada, provided a scientific review of the manuscript. We are very grateful to Pamela Cheers, Canadian Forest Service and to Alix Rive for reviewing the manuscript and providing helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.geoderma.2014.06.032.

References

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A.B. (Eds.), 2014. GlobalSoilMap: Basis of the global spatial soil information system. CRC Press.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., Hall, R.J., 2014. Mapping attributes of Canada's forests at moderate resolution through *k*NN and MODIS imagery. Can. J. For. Res. 44, 521–532.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168, 21–33.

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175–185.

Bernier, P.Y., Daigle, G., Rivest, L.P., Ung, C.H., Labbé, F., Bergeron, C., Patry, A., 2010. From plots to landscape: a *k*NN based method for estimating stand-level merchantable volume in the Province of Québec, Canada. For. Chron. 86, 461–468.

Boettinger, J.L., 2010. Environmental covariates for digital soil mapping in the western USA. Digital Soil Mapping: Bridging Research, Production, Environmental Application and Operation, vol. 2. Springer, Netherland.

Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16, 199–215.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modeling and legacy data. Geoderma 103, 79–94.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111, 21–44.

Ecological Stratification Working Group, 1996. A National Ecological Framework for Canada. Agriculture and Agri-Foods Canada, Research Branch, Centre for Land and Biological Resources Research and Environment Canada, State of the Environment Directorate, Ottawa (125 pp.).

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–449.

Franco-Lopez, H., Ek, A.R., Bauer, M.E., 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sens. Environ. 77, 251–274.

Geng, X., Fraser, W., VandenBygaart, B., Smith, S., Waddell, A., Jiao, Y., Patterson, G., 2010. Toward digital soil mapping in Canada: existing soil survey data and related expert knowledge. Digital Soil Mapping. Springer, Netherlands, pp. 325–335.

Gillis, M.D., Omule, A.Y., Brierley, T., 2005. Monitoring Canada's forests: the National Forest Inventory. For. Chron. 81, 214–221.

Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2014. Lithology as a powerful covariate in digital soil mapping. GlobalSoilMap: Basis of the Global Spatial Soil Information System. Taylor & Francis Group, London, UK, pp. 433–439.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Sci. Soc. Am. J. 75, 1201–1213.

Guindon, L., Bernier, P.Y., Beaudoin, A., Pouliot, D., Villemaire, P., Hall, R.J., Latifovic, R., St-Amant, R., 2014. Annual mapping of severe forest disturbances across Canada's forests using 250 m MODIS imagery from 2000 to 2011. Can. J. For. Res. (submitted for publication).

Hahm, W.J., Riebe, C.S., Lukens, C.E., Araki, S., 2014. Bedrock composition regulates mountain ecosystems and landscape evolution. PNAS. http://dx.doi.org/10.1073/pnas.1315667111.

Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (Eds.), 2008. Digital Soil Mapping With Limited Data. Springer, Dordrecht (445 pp.).

Hazlett, P.W., Morris, D.M., Fleming, R.L., 2014. Effects of biomass removals on site carbon and nutrients and jack pine growth in boreal forests. Soil Sci. Soc. Am. J. http://dx.doi.org/10.2136/sssaj2013.08.0372nafsc.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214, 141–154.

Jackson, R.B., Canadell, J., Ehleringer, J.R., Mooney, H.A., Sala, O.E., Schulze, E.D., 1996. A global analysis of root distributions for terrestrial biomes. Oecologia 108, 389–411.

Laamrani, A., Valeria, O., Fenton, N., Bergeron, Y., Cheng, L.Z., 2014. The role of mineral soil topography on the spatial distribution of organic layer thickness in a paludified boreal landscape. Geoderma. http://dx.doi.org/10.1016/j.geoderma.2014.01.003.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133, 90–99.

Ladd, B., Laffan, S.W., Amelung, W., Peri, P.L., Silva, L.C.R., Gervassi, P., Bonser, S.P., Navall, M., Sheil, D., 2013. Estimates of soil carbon concentration in tropical and temperate forest and woodland from available GIS data on three continents. Glob. Ecol. Biogeogr. 22, 461–469.

Lagacherie, P., McBratney, A.B., 2006. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Dev. Soil Sci. 31, 3–22.

Magnussen, S., Tomppo, E., McRoberts, R.E., 2010. A model-assisted k-nearest neighbour approach to remove extrapolation bias. Scand. J. For. Res. 25, 174–184.

McBratney, A.B., Mendonça-Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McKenney, D.W., Hutchinson, M.F., Papadopol, P., Lawrence, K., Pedlar, J., Campbell, K., Milewska, E., Hopkinson, R., Price, D., Owen, T., 2011. Customized spatial climate models for North America. Bull. Am. Meteorol. Soc. 92, 1612–1622.

McRoberts, R.E., 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. For. Ecol. Manag. 272, 3–12.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sens. Environ. 111, 466–480.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. Digital Soil Mapping. Springer, Netherlands, pp. 429–436.

Ministère des Ressources naturelles du Québec (MRN), 2013. Carte des dépôts de surface au 1:40,000. Direction des inventaires forestiers, Québec.

Mount, D.M., Arya, S., 2010. ANN: A Library for Approximate Nearest Neighbor Searching. http://www.cs.umd.edu/~mount/ANN/.

Nemes, A., Wösten, J.H.M., Lilly, A., Oude Voshaar, J.H., 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. Geoderma 90, 187–202.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70, 327–336.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91–100.

Post, W.M., Emanuel, W.R., Zinke, P.J., Stangenberger, A.G., 1982. Soil carbon pools and world life zones. Nature 298, 156–159.

Pouliot, D., Latifovic, R., Fernandes, R., Olthof, I., 2009. Evaluation of annual forest disturbance monitoring using a static decision tree approach and 250 m MODIS data. Remote Sens. Environ. 113, 1749–1759.

Scheuber, M., 2010. Potentials and limits of the k-nearest-neighbour method for regionalising sample-based data in forestry. Eur. J. For. Res. 129, 825–832.

Schut, P., Smith, S., et al., 2011. Soil landscapes of Canada: building a national framework for environmental information. Geomatica 65, 293–309.

Seibert, J., Stendahl, J., Sorensen, R., 2007. Topographical influences on soil properties in boreal forests. Geoderma 141, 139–148.

Sequeira, C.H., Wills, S.A., Seybold, C.A., West, L.T., 2014. Predicting soil bulk density for incomplete databases. Geoderma 213, 64–73.

Soil Landscapes of Canada (SLC) v.2.2, 1996. Centre for Land and Biological Resources Research. Soil Landscapes of Canada, v.2.2Research Branch, Agriculture and Agri-Food Canada, Ottawa (Available at http://sis.agr.gc.ca/cansis/nsdb/slc/v2.2/index.html).

Soil Landscapes of Canada (SLC) v.3.2, 2010. Centre for Land and Biological Resources Research. Soil Landscapes of Canada, v.3.2Research Branch, Agriculture and Agri-Food Canada, Ottawa (Available at http://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/index.html).

Suominen, L., Ruokolainen, K., Tuomisto, H., Llerena, N., Higgins, M.A., 2013. Predicting soil properties from floristic composition in western Amazonian rain forests: performance of k-nearest neighbour estimation and weighted averaging calibration. J. Appl. Ecol. 50, 1441–1449.

Thiffault, E., Paré, D., Brais, S., Titus, B.D., 2010. Intensive biomass removals and site productivity in Canada: a review of relevant issues. The forestry chronicle 86, 36–42.

Thiffault, E., Paré, D., Guindon, L., Beaudoin, A., Brais, S., Leduc, A., Michel, J.-P., 2013. Assessing forest soil base cation status and availability using lake and stream sediment geochemistry: a case study in Quebec (Canada). Geoderma 211, 39–50.

Thiffault, E., Barrette, J., Paré, D., Titus, B.D., Keys, K., Morris, D.M., Hope, G., 2014. Developing and validating indicators of site suitability for forest harvesting residue removal. Ecol. Indic. 43, 1–18.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., Katila, M., 2008. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sensing of Environment 112, 1982–1999.

Veillette, J.J., 1994. Evolution and paleohydrology of glacial Lakes Barlow and Ojibway. Quat. Sci. Rev. 13, 945–971.

Welsch, D.L., Kroll, C.N., McDonnell, J.J., Burns, D.A., 2001. Topographic controls on the chemistry of subsurface stormflow. Hydrol. Process. 15, 1925–1938.

Wilson, B.T., Lister, A.J., Riemann, R.I., 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. For. Ecol. Manag. 271, 182–198.

Xu, Y., Dickson, B.G., Hampton, H.M., Sisk, T.D., Palumbo, J.A., Prather, J.W., 2009. Effects of mismatches of scale and location between predictor and response variables on forest structure mapping. Photogramm. Eng. Remote Sens. 75, 313–322.