



HAL
open science

Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France

Songchao Chen, Anne C Richer-De-Forges, Vera Leatitia Mulder, Guillaume Martelet, Thomas Loiseau, Sébastien Lehmann, Dominique Arrouays

► To cite this version:

Songchao Chen, Anne C Richer-De-Forges, Vera Leatitia Mulder, Guillaume Martelet, Thomas Loiseau, et al.. Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France. CATENA, 2021, 198, pp.105062. 10.1016/j.catena.2020.105062 . hal-03467117

HAL Id: hal-03467117

<https://hal.inrae.fr/hal-03467117>

Submitted on 3 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Title: Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France

Authors:

Songchao Chen^a. songchao.chen@inrae.fr

Anne C. Richer-de-Forges^a. anne.richer-de-forges@inrae.fr

Vera Leatitia Mulder^b. titia.mulder@wur.nl

Guillaume Martelet^c. g.martelet@brgm.fr

Thomas Loiseau^a. thomas.loiseau@inrae.fr

Sébastien Lehmann^a. sebastien.lehmann@inrae.fr

Dominique Arrouays^a. dominique.arrouays@inrae.fr

Affiliations:

^a INRAE, Unité InfoSol, 45075 Orléans, France

^b Soil Geography and Landscape Group, Wageningen University, PO Box 47 6700 AA Wageningen, The Netherlands

^c BRGM, UMR 7327, 45060 Orléans, France

Corresponding author:

Songchao Chen: songchao.chen@inrae.fr

Postal address: INRAE, Unité InfoSol, 2163 Avenue de la Pomme de Pin, CS 40001 Ardon, 45075 Orléans, France

1 **Abstract:**

2 Soil thickness (ST) plays an important role in regulating soil processes,
3 vegetation growth and land suitability. Therefore, it has been listed as one of twelve
4 basic soil properties to be delivered in *GlobalSoilMap* project. However, ST prediction
5 has been reported with poor performance in previous studies. Our case study is
6 located in the intensive agriculture Beauce area, central France. In this region, the ST
7 mainly depends on the thickness of loess (TOL) deposits over a calcareous bedrock.
8 We attempted to test the TOL prediction by coupling a large soil dataset (10978
9 sampling sites) and 117 environmental covariates. After variable selection by
10 recursive feature elimination, quantile regression forests (QRF) was employed for
11 spatial modelling, as it was able to directly provide the 90% prediction intervals (PIs).
12 Averaging a total of 50 models, generated by repeated stratified random sampling,
13 showed a substantial model performance with mean R^2 of 0.33, RMSE of 30.48 cm
14 and bias of -1.20 cm. The prediction interval coverage percentage showed that 86.70%
15 of the validation samples fall within the predefined 90% PIs, which also indicated the
16 prediction uncertainty produced by QRF was reasonable. The relative variable
17 importance indicated the importance of airborne gamma-ray radiometric data and
18 Sentinel 2 products in TOL prediction. The produced TOL map with 90% PIs makes
19 sense from a soil science and physiographic point of view. The final product can
20 guide evidence-based decision making for agricultural land management, especially
21 for irrigation in our case study.

22 1. Introduction

23 Soil thickness (ST) is an important soil property due to its influence as a
24 controlling factor of numerous surface and subsurface soil processes. Through its
25 influence on the plant rootable depth (Leenaars *et al.*, 2018), ST is a major
26 controlling factor of vegetation growth and land suitability, and is a key variable to
27 estimate available water capacity (AWC). As a consequence, ST has been retained
28 as a mandatory soil attribute to be delivered in *GlobalSoilMap* products (Arrouays *et*
29 *al.*, 2014). Previously, some attempts have been made to map ST at national,
30 continental or global levels (e.g., Grundy *et al.*, 2015; Lacoste *et al.*, 2016; Mulder *et*
31 *al.*, 2016; Hengl *et al.*, 2017; Chen *et al.*, 2019). Grundy *et al.* (2015) mapped ST in
32 Australia using about 300,000 points of observations and environmental covariates
33 as inputs for a Cubist model. Lacoste *et al.* (2016) tested three digital soil mapping
34 (DSM) approaches, based on regression tree modelling, gradient boosting modelling,
35 and multi-resolution kriging for a dataset of ca 14,000 observations in France. Hengl
36 *et al.* (2017) mapped ST at a global scale. Predictions were based on ca. 150,000
37 soil profiles used for training and a stack of 158 covariates which were used to fit an
38 ensemble of machine learning methods—random forest and gradient boosting and/or
39 multinomial logistic regression. Mulder *et al.* (2016) used Cubist predictions for
40 mainland France using ca 16,000 observation points and a set of 20 spatially
41 exhaustive covariates. Chen *et al.* (2019) further demonstrated how right-censored
42 data can be accounted for in the ST modelling of mainland France. Using random
43 survival forest, the probability of exceeding a given depth was modelled using freely
44 available spatial data representing the main soil-forming factors. However, most of
45 these results gave rather poor prediction performances compared to other soil
46 properties such as soil organic carbon or clay content and pH (e.g., Mulder *et al.*,
47 2016; Hengl *et al.*, 2017). In many cases, ST prediction proved to be hampered either
48 1) by the lack of data measurements (Leenaars *et al.*, 2018) or 2) by the fact that the
49 collected ST data is often right-censored data (i.e. the observed ST is less than true
50 ST, Chen *et al.*, 2019), or 3) because ST has a high short-range spatial variability in
51 specific pedological contexts (e.g., Bourennane *et al.*, 1996; Lacoste *et al.*, 2016).
52 Moreover, most of the examples taken from the literature were produced using digital
53 soil mapping (DSM, McBratney *et al.*, 2003) and in some cases, one may expect that
54 no relevant covariate was available to improve the performance of the predictions.

55 For example, while topography-related covariates such as elevation or slope often
56 explain large part of the variability of soils in a given areas, these covariates might
57 not improve the performance of predictions in flat areas.

58 Mapping ST enables several applications among which are agronomy and
59 agricultural practices (plant rootable depth, drainage, irrigation), crop growth
60 modelling, geotechnical engineering, water balance modelling at catchment to global
61 scale, and Quaternary science studies.

62 The Beauce area, located in central France is a limestone plateau irregularly
63 covered with Quaternary loessic silt (Macaire, 1971; Lorain, 1973; M nillet, 1974).
64 Soil classification varies from Luvisols to Calcic and Calcic Cambisols. The region
65 is rather flat, and has a noticeable proportion of rather thin soils. Intensive agriculture
66 in this region often utilizes irrigation and most of the surface area is occupied by
67 cereal crops (mainly maize and wheat) and sugar beet. Also, the Beauce area is
68 home of the largest aquifer of France in the underlying calcareous rock. Upper
69 horizons were affected by peri-glacial winds that redistributed loess deposits
70 (Macaire, 1971; Bourennane *et al.*, 1996), resulting in a rather homogeneous
71 particle-size distribution of the fine earth (i.e., silt, clay, and sand). Therefore, most of
72 the soils consist now of silt, silt loam or silty clay loam layers derived from this aeolian
73 deposit developed on a lacustrine limestone substrate. In general, the illuviation
74 process occurred when the thickness of loess was the largest, resulting in less clayey
75 topsoil textures.

76 The available water capacity (AWC) is the maximum amount of available water
77 that can be stored for crop growth, therefore it is an important soil information for
78 agricultural management. Therefore, the thickness of loess (TOL) deposit is also one
79 of the primary factors influencing the calculation of the soil AWC. Although Tetegan
80 *et al.* (2015) demonstrated that the percentage of rock fragments was also one of the
81 controlling factors of AWC in this region. Overall, in this region, irrigation
82 management is of utmost importance in order to maintain crop yields, while
83 preserving the underlying water table and water quality. Knowing the TOL is essential
84 for determining a water balance and for piloting irrigation management. In terms of
85 agronomy and environment, the TOL is a determining factor (Nicoullaud *et al.*, 1995;
86 Ould-Mohamed *et al.*, 1997). Therefore, the TOL should be known accurately and
87 cheaply over the study area. Several traditional soil maps have been produced in this

88 region, with scales ranging from 1:50,000 to 1:250,000, resulting in a various density
89 of point scale soil information.

90 The objective of this study is to assess to which extent using this legacy data and
91 environmental covariates (from existing geological maps, digital elevation model
92 derivatives, airborne gamma-ray radiometry, and remote sensing data) in a DSM
93 model allows to reach acceptable performances for TOL prediction. In this study, we
94 decided to model the TOL up to a depth of 120 cm using Quantile Regression
95 Forests (QRF) because the TOL was deemed useful for agricultural practices. Maize
96 cropping is especially of interest because it is known for its high water requirement
97 (Doorenbos *et al.*, 1978) and thus typically requires the largest amounts of irrigation.
98 The average rooting depth of maize is equal to 120 cm (British Standards Institution,
99 1988; Tetegan *et al.*, 2015). Therefore, we only mapped the TOL up to a depth of
100 120 cm, as there is no difference of soil water management between soils with a TOL
101 deeper than 120 cm and soils with a TOL of 120 cm.

102

103 **2. Material and methods**

104 **2.1. Study area**

105 This study was conducted in the Beauce area located at the middle Loire
106 catchment, central France (Figure 1). It covers a total area of 4835 km², of which
107 agriculture is the dominant land use (88.5%, Inglada *et al.*, 2017). It has a
108 continental-oceanic climate with a mean annual temperature of 11.5°C and a mean
109 annual rainfall of 700 mm (Paroissien *et al.*, 2014). Most of the soils in this study area
110 are developed from periglacial loess deposits which covered a limestone bedrock.
111 Cambisol (48.3%) and Luvisol (25.6%) are the major soil groups observed in this
112 region (IUSS Working Group WRB, 2006). At the southern border of the Beauce
113 region, some other soil groups (not developed from loess) are observed.

114 **2.2. Soil data**

115 We used available soil data from the French Soil Inventory Program (IGCS). The
116 thickness of loess derived horizons (TOL) was determined by several criteria: 1)
117 digging soil pits down to the calcareous material and 2) by auger borings. The
118 presence of small rock fragments could in some case lead to an underestimation of
119 TOL done by augering. Therefore, if a TOL of 120 cm was not reached, two other

120 augerings were made randomly 0.5 m apart from the first one and the maximum TOL
121 reached was recorded. TOL should have a texture of silt, silt loam or silty clay loam
122 (Bertran *et al.*, 2016; Borderie *et al.*, 2017). The deeper the TOL the more the
123 illuviation processes are pronounced and the lighter the topsoil texture.

124 In total, 10978 sites were used in this study to map the TOL up to a depth of 120
125 cm. The TOL for sites with a TOL deeper than 120 cm (n=14) was set to 120 cm
126 before modelling to eliminate the effect of extreme values in modelling.

127 **2.3. Environmental covariates**

128 The environmental covariates used in this study and their data sources are listed
129 in Table 1. These covariates provide information on the environmental factors
130 assumingly controlling TOL, based on the Scorpan conceptual model (McBratney *et al.*
131 *et al.*, 2003). For illustrative purposes, several covariates are shown in Figure 2.

132 *2.3.1. Relief*

133 The Digital Elevation Model of mainland France was derived from BD TOPO 3 of
134 the French National Geographical Institute (IGN, 2011), at 25 m resolution. SAGA
135 GIS (Conrad *et al.*, 2015) was used to calculate its derivatives (relief factors),
136 including channel network base level (CNBL), multiresolution index of valley bottom
137 flatness (MrVBF), plan curvature (PIC), profile curvature (PrC), slope (SI), slope
138 position (SIP), slope length (SIL), terrain wetness index (TWI), valley depth (VD), and
139 vertical distance to channel network (VDCN). As the relief factor at neighbouring
140 locations is able to provide additional useful information in modelling soil patterns
141 (McBratney *et al.*, 2003), some previous studies investigated the potential of
142 incorporating local neighbourhood information into the training pixels, using
143 convolution filtering operations (e.g., Grinand *et al.*, 2008; Loiseau *et al.*, 2019).
144 Filtering can be achieved by passing a moving window over the variable to calculate
145 a value of the processing cell (central pixel) using the values of its neighbouring cells.
146 In this study, we used mean convolution circular windows to calculate the focal
147 means for these relief factors with radius at 200, 500 and 1000 m (Grinand *et al.*,
148 2008), which resulted in three raster layers derived from each original relief factor (25
149 m).

150 *2.3.2. Soil*

151 The soil type information were extracted from the French national soil type map
152 at 1:1 M scale (King *et al.*, 1994). The soil types in this study area were mainly
153 Cambisols and Luvisols. However, some Podzols, Gleysols, Fluvisols, Arenosols and
154 Vertisols were rarely present, mainly at the southern border of the region.

155 *2.3.3. Parent material*

156 The map of parent material was extracted from the French national parent
157 material map (King *et al.*, 1994). Undifferentiated alluvial deposits, calcareous rocks,
158 clayey materials, sandy materials and loamy materials are the main parent materials
159 in the study area. Note that the loamy materials are nearly always located over
160 underlying calcareous rocks.

161 The gamma radiometric data, including Potassium (K), Thorium (Th) and
162 Uranium (U), and total count (TC), was derived from an airborne high-resolution
163 magnetic and radiometric survey over the Région Centre, flown by Terraquest Ltd,
164 Canada, under the supervision of BRGM between 2008 and 2009 (Martelet *et al.*,
165 2014). The line-spacing of the survey was 1 km and, along the flight lines the
166 footprint of each gamma radiometric measurement was an ellipse of 150 × 250 m²;
167 accordingly the data were interpolated on 250 m grids using a standard minimum
168 curvature interpolation.

169 *2.3.4. Organisms*

170 A land use map was extracted from the French land use map, which was
171 produced from Sentinel 2 data at 10 m resolution, for year 2016 (Inglada *et al.*, 2017).
172 This land use map was aggregated to 25 m resolution by majority sampling and the
173 proportions (0~100%) of the nine main land-use classes within each 25×25 m pixel
174 (which contained 6 10×10 pixels) were also included as covariates.

175 The monthly normalized difference vegetation index (NDVI) from the MODIS
176 (MCD43A4 16-day Version 6) in 500 m resolution and the PROBA-V 10-day product
177 level 2B TOC (Copernicus, 2016) in 300 m resolution were used in this study. These
178 24 monthly NDVI data in 2003 (extreme warm and dry year) and 2016 (normal year)
179 were collected and reduced into the first three principal components by principal
180 component analysis to eliminate their multicollinearity. For more details, we refer to
181 Loiseau *et al.* (2019).

182 We also included 42 covariates related to Sentinel 2 bands (year of 2016 to 2017)
183 and indices, which were produced in an earlier study from Loiseau *et al.* (2019) for
184 mainland France at 90 m resolution. The Sentinel 2 data were processed to Level-2A
185 (atmospheric and topographic corrections) by the French National Centre for Space
186 Studies (Hagolle *et al.*, 2015). These covariates included 10 Sentinel 2 bands (2, 3, 4,
187 5, 6, 7, 8, 8A, 11 and 12), 11 spectral indices (brightness index, saturation index, hue
188 index, coloration index, redness Index, carbonate index, ferrous iron, clay index,
189 normalized difference 1, normalized difference 2 and grain size index) and their focal
190 means determined by a low-pass filter with an average within a 2×2 km window. For
191 more details, we refer to Loiseau *et al.* (2019).

192 2.3.5. Position

193 The coordinates, i.e., latitude and longitude (extracted for each recorded
194 sampling site), were used in modelling. In addition, 10 oblique geographic
195 coordinates were calculated at angles of 15°, 30°, 45°, 60°, 75°, 105°, 120°, 135°,
196 150° and 165°. The oblique coordinate (OC) at an angle of θ can be calculated as
197 below (Møller *et al.*, 2019):

$$198 \text{ OC} = \sqrt{X^2 + Y^2} \times \cos(\theta - \tan^{-1}(\frac{Y}{X})) \quad (1)$$

199 where X and Y are the latitude and longitude.

200 Note that when θ is 0° or 90°, the oblique coordinate equals to latitude or
201 longitude.

202 2.3.6. Harmonization of environmental covariates

203 The environmental covariates had different resolutions and scales, we therefore
204 harmonized them at 25 m resolution using nearest neighbour interpolation for spatial
205 predictive modelling and mapping at non-visited locations.

206 2.4. Variable selection using recursive feature elimination

207 Considering the large set of environmental covariates ($n=117$), variable selection
208 was applied by recursive feature elimination (Kunn, 2020) prior to fitting the spatial
209 predictive model. The recursive feature elimination (incorporating resampling) adopts
210 a backwards selection, which includes several steps: (1) split data into training and
211 test set by resampling (i.e., k -fold cross-validation); (2) train the model on the training
212 set using all predictors, calculate the model performance on the test set, and rank

213 predictors using their model importance; (3) for each predictor subset size S_i ($i=1,$
214 $2, \dots, s$), train the model on the training set using the S_i most important predictors,
215 and calculate the model performance on the test set; (4) compare the model
216 performance profile over the S_i on the test set, and determine the optimal number of
217 predictors.

218 To select the important covariates and improve the mapping efficiency, the
219 recursive feature elimination was performed on the whole data using *rfe* function in
220 *caret* package (Kunn, 2020) in R (R Core Team, 2019). The model was set to
221 Random Forest (default values with tree number of 500 and $mtry$ of $p/3$ where p is
222 the size of predictors) using 5-fold cross-validation. Seven predictor subset sizes (5,
223 10, 15, 20, 40, 60, 80 and 100) were tested and the model performance indicated
224 that 80 variables (Table 2) were optimal and then used for later modelling.

225 **2.5. Spatial predictive modelling and model performance evaluation**

226 Quantile Regression Forest (QRF, Meinshausen, 2006) has been growingly used
227 in DSM for delivering soil information as it is able to provide uncertainty estimates
228 straightforwardly with a fair model performance (e.g., Vaysse and Lagacherie, 2017;
229 Lombardo *et al.*, 2018; Loiseau *et al.*, 2019). Therefore, QRF was used for modelling
230 TOL in this study.

231 Since QRF is an extension of Random Forest (RF, Breiman, 2001), we start with
232 RF. Assume X and Y are the predictor variables and responses, for regression, RF
233 generates a large number (b) of bootstrap trees by using m training samples $(X_i, Y_i),$
234 $i=1, \dots, m$. Here, bootstrap refers to repeated (b times) selection of a random sample
235 with replacement of the training samples. For each node in a bootstrap tree, a
236 random subset of the predictor variables is used for split-point selection. The
237 prediction of a bootstrap tree for a new sample $D=X_d$ is the conditional mean
238 estimate (\hat{X}) of Y , which can be represented by:

$$239 \hat{X} = \sum_{i=1}^m w_i Y_i \quad (2)$$

240 where w_i is the weight of the sample (X_i, Y_i) in the same leaf of the bootstrap tree.

241 The final prediction of the new sample D is approximated by the mean predictions of
242 b bootstrap trees.

243 Apart from the conditional mean estimate in RF, QRF also uses the weighted
244 samples to derive a conditional distribution. This distribution function is able to
245 provide the probability of Y being lower than a given percentile and thus to calculate
246 the prediction intervals. For more details about the constructions of the conditional
247 distribution, we refer to Meinshausen (2006).

248 We used the *quantregForest* package (Meinshausen, 2017) in R (R Core Team,
249 2019) for implementing QRF to derive the median prediction and 90% prediction
250 intervals (90% PIs, 5th and 95th quantiles). The default number of tree (*ntree*=500)
251 and minimum size of terminal nodes (*nodesize*=5) were used for QRF, and the
252 number of variables randomly sampled as candidates at each split (*mtry*) was
253 optimized in the *caret* package (Kunn, 2020) by 5-fold cross-validation in R (R Core
254 Team, 2019). The variable importance was determined by the increased mean
255 square error (IncMSE, in %) between the model excluding and including a given
256 variable, and this information was integrated in QRF model. In our case, the variable
257 importance was calculated by the average of 50 repeated models.

258 Considering the highly varying soil sampling density (Brus *et al.*, 2011), we
259 divided the study area into 20 compact equal area geographical strata (Figure 3)
260 using the *spsosa* package (Walvoort *et al.*, 2020) in R (R Core Team, 2019), and
261 performed stratified random sampling (5 sites for each strata) for selecting the
262 validation set. It resulted in a set of 10878 sites for model calibration and 100 sites for
263 model validation. To derive a robust result, we repeated this procedure 50 times and
264 took the average as the final model performance.

265 Four indicators were used to evaluate the model performance in validation set: (1)
266 modelling efficiency (R^2); (2) root mean square error (RMSE); (3) bias; (4) prediction
267 interval coverage percentage (PICP), which describes the percentage of the
268 observed TOL falls within the estimated upper and lower 90% PIs.

$$269 \quad R^2 = 1 - \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - y_i)^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3)$$

270
$$\text{RMSE} = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

271
$$\text{Bias} = \sum_{i=1}^n \frac{(\hat{y}_i - y_i)}{n} \quad (5)$$

272 where n is the sample size of observations, y_i and \hat{y}_i are observed value and
 273 predicted value for sample i , and \bar{y} is the average of observed values.

274 In addition, we also reported the model performance by the internal validation
 275 using out-of-bag data (around 34% of data that is not used for growing each tree) in
 276 QRF. The final TOL map and its 90% PIs were produced by QRF fitted using all the
 277 sampling sites.

278

279 **3. Results and discussion**

280 **3.1. Summary of TOL in the Beauce area**

281 Table 3 displays the statistics of the TOL in the Beauce area. Among 10978 sites,
 282 TOL ranged from 0 to 120 cm, with a mean and median TOL at 36.46 cm and 30 cm,
 283 respectively. A skewness of 0.66 (larger than 0.5) indicated the data were slightly
 284 positively skewed while a kurtosis less than 3 (2.85) showed that the data were light-
 285 tailed. Though the log transformation, i.e. $\log(\text{TOL}+a)$, is able to convert soil data to
 286 normal distribution (a skewness less than 0.5 and a kurtosis close to 3), it did not
 287 improve the model performance. Therefore, instead of data transformation, we used
 288 the original TOL data for spatial modelling in this study.

289 Figure 4 presents the TOL located in 20 compact equal area geographical strata.
 290 It showed a large difference of number of sampling sites among these 20
 291 geographical strata, ranging from 22 to 1351. These geographical strata with high
 292 median TOL (>60 cm) had much less sampling sites (51 to 201), and this is the main
 293 reason for evaluating the model performance by random stratified sampling.

294 Figure 5 shows the Pearson correlations coefficients between TOL and top 30
 295 environmental covariates. Elevation and its focal mean derivatives had the highest
 296 positive correlations ($r > 0.3$). Other positively correlated covariates were related to

297 oblique coordinates, channel network base level (CNBL), graphic coordinates,
298 gamma-ray radiometry and two Sentinel 2 indexes (grain size index and clay index).
299 Negative correlations with TOL were found with valley depth, Sentinel 2 bands and its
300 indexes. Overall, the correlations between TOL and covariates were found to be
301 rather low ($|r| < 0.35$).

302 **3.2. Model performance of Quantile Random Forest**

303 The mean R^2 , RMSE and bias from the internal validation using the out-of-bag
304 data in QRF were 0.31, 26.88 cm and 0.23 cm (data not shown). Figure 6 indicates
305 the model performance after we repeated 50 times the validation procedure using
306 QRF. The mean R^2 and RMSE were 0.33 and 30.48 cm respectively. The mean bias
307 of -1.20 cm indicated that the prediction was almost un-biased for 50 repeats. The
308 mean PICP indicated that on average 86.7% of the validation samples fall within the
309 defined 90% PIs, therefore the uncertainty estimates from the QRF model was valid
310 for non-visited locations.

311 As shown in Table 4, the global soil thickness (ST) products (Hengl *et al.*, 2017;
312 Shangguan *et al.*, 2017) had better model performance than those at national or
313 regional scale (Guerrero *et al.*, 2014; Kidd *et al.*, 2015; Vaysse and Lagacherie, 2015;
314 Lacoste *et al.*, 2016; Mulder *et al.*, 2016; Zhang *et al.*, 2018). This may be attributed
315 to the fact that global ST products include a substantial proportion of very thin soils
316 (i.e., soils prone to severe erosion) and of very thick ones (i.e., Arenosols in desert
317 dunes, Shangguan *et al.*, 2017). There was no large difference of model performance
318 between national and regional products, even if we used nearly 11000 sampling sites
319 in this study. This is because the TOL is highly variable at short distances. By
320 incorporating a large exhaustive set of environmental covariates, however, the map
321 produced in this study performed slightly better than almost all the previous studies at
322 regional and national scales.

323 The large range between upper and lower limits of 90% PIs for R^2 and RMSE
324 indicated the randomness involved in data split brought a large amount of uncertainty
325 in model evaluation. Therefore, instead of a single time data split, repeated random
326 (stratified) sampling adopted in this study would provide more robust estimates for
327 the model performance so as to avoid under- or over- optimistic decision making in
328 management of soil resources.

329 **3.3. Variable importance of environmental covariates**

330 Figure 7 displays the top 30 environmental covariates in QRF model calculated
331 as the average of 50 repeats. It indicated that the gamma radiometric data (U, Th and
332 TC) and hue index (focal mean) calculated from Sentinel 2 images were the most
333 important environmental covariates in modelling TOL in the study area. They were
334 followed by longitude, NDVI PC1, slope position (with a radius of 1000 m), TWI (with
335 a radius of 1000 m), normalized difference (focal mean) and grain size index (focal
336 mean), representing position, organisms and relief factors in Scorpan conceptual
337 model. For many relief (e.g., slope position, TWI, VDCN, curvature, valley depth,
338 CNBL, elevation, slope) and organisms (e.g., hue index, normalized difference, grain
339 size index, ferrous iron) factors, their derivatives calculated from neighbouring
340 information performed better than original covariates. Spatial position, i.e., latitude
341 and longitude, were identified important in Figure 7 while oblique coordinates were
342 not listed among the top 30 covariates.

343 Interestingly, the variable importance in the QRF model was not in line with the
344 correlations between TOL and covariates (see Figure 5). This may be due to the fact
345 that the relations between TOL and covariates are not linear. If the relationships were
346 linear then the most important covariates should have been those with highest $|r|$
347 which is not the case in this study. Another reason may be that the importance of
348 covariates results also from interactions between them, that are not visible using
349 Pearson correlations but that are taken into account in QRF model.

350 Our results indicate a high importance of airborne gamma radiometric data in
351 TOL modelling as they can capture soil information relevant to soil texture and to the
352 presence of the calcareous rock at low depth. Indeed, the substrate of part of the
353 study area (composed carbonates) is completely different from the TOL and it has
354 been shown that calcium mitigates surface gamma-spectrometric signatures because
355 it has a poor gamma-spectrometric response (Martelet *et al.*, 2013). Therefore, it is
356 not surprising that gamma radiometric data plays an important role, especially for
357 predicting thin TOL. Also, the large plateaus with deep TOL in the northern part are
358 depleted in K (see Figure 2). This is because soils with large TOL were prone to
359 illuviation, resulting in lower clay content in topsoil. So in this case, it is an indirect
360 relationship with TOL. Our results also confirm the contribution of neighbouring
361 information (e.g, focal hue index, slope position 1000m, TWI 1000m, focal normalized

362 difference) of relief and organism factors in spatial modelling of TOL, which
363 implicates the multi-scale influence of covariates on soil properties. Concerning slope
364 and TWI, the importance of this neighbouring information may be due to the gradient
365 of TOL that shows that very large flat plateaus (mainly in the north) are characterized
366 by a deeper TOL. These derivatives likely performed well because it is not the same
367 geomorphological context if you have a flat location inside a very large flat plateau
368 than if you have locally flat “pixels” in a region where the relief is more accentuated,
369 such for instance in the southwest (Behrens *et al.*, 2019). Other studies also have
370 shown the potential of multi-scale covariates derivatives in improving model
371 performance in DSM (Behrens *et al.*, 2018b, 2019). Compared to simple convolution
372 approach (focal mean), wavelet transforms, empirical mode decomposition, and the
373 Gaussian scale space may even better represent the multi-scale information of
374 environmental covariates (Behrens *et al.*, 2010, 2018a, 2018b; Biswas *et al.*, 2013a,
375 2013b; Zhou *et al.*, 2016; Huang *et al.*, 2017; Zhao *et al.*, 2018; Liang *et al.*, 2019) so
376 as to improve model performance in DSM. The Sentinel 2 spectral bands may not
377 always provide direct information related to soil, while a great potential has been
378 shown from its derived indicators (e.g., NDVI, hue index, normalized difference,
379 ferrous iron, and grain size index) in this study. Considering its high spatial and
380 temporal resolution, Sentinel 2 has a great potential in delivering useful information of
381 soil surface for DSM across scales (Gholizadeh *et al.*, 2018; Castaldi *et al.*, 2019;
382 Loiseau *et al.*, 2019; Vaudour *et al.*, 2019). Some of the Sentinel 2 data we used
383 come from a mosaic of images of bare soils built by Loiseau *et al.* (2019). Therefore,
384 these Sentinel 2 data provide direct information on soil colour which may reflect thin
385 TOL or absence of TOL through the presence of white calcareous rocks at the
386 surface. They may indirectly reflect also texture through bright colours due to slaking
387 that occurs mainly on very loamy topsoil soils which correspond to the deepest TOL
388 where Luvisols have developed. The land use map produced by Sentinel 2 was not
389 among the top 30 environmental covariates as it may be masked by the NDVI data
390 due to their correlation or NDVI better explains the spatial variability than land use
391 map. Therefore, relative importance of environmental covariates should be taken with
392 caution as high contributing covariates can inadvertently bear part of the contribution
393 of the less contributing covariates (Chen *et al.*, 2018).

394 **3.4. Maps of thickness of loess and its 90% prediction intervals**

395 Figure 8 presents the spatial distribution of TOL and its lower and upper limits of
396 90% PIs. It displays the general increasing thickness of loess soils from south-west to
397 north in the study area. Very shallow loess (<10 cm) was found in south-west of the
398 study area, and very deep loess (>100 cm) was mainly found in the northern part.
399 Highest TOL were mainly located in rather flat areas located on high elevation
400 plateaus, while shallow loess was mainly located at lower elevations and in more
401 dissected relief, especially in the vicinity of small valleys. Note that there is a border
402 effect from the south-west to the west of the region. This border effect corresponds to
403 the outcropping limit of the TOL, where sandy or clayey materials locally overlay the
404 calcareous. The regions with thin soil (<10 cm with a lot of outcrops of the calcareous)
405 correspond to the areas with black gamma-ray radiometry patterns matching on
406 steep slopes around the drainage lines (mostly rivers).

407 The maps of lower and upper limits of 90% PIs clearly show different spatial
408 structures. On the northern part with the highest elevations and high mean TOL, the
409 95th percentile is equal or deeper than 1.2 m, which means that high TOL are largely
410 dominant in these plateaus. On the contrary the extreme southern part of the region
411 exhibits TOL that rarely exceed 0.6 m. Moreover, except for some very local areas
412 having a high mean TOL, the 5th percentile map suggests that the upper calcareous
413 surface is undulating at very short distances and that local calcareous outcrops may
414 be found in nearly all the southern part of the study area. The wind direction of loess
415 deposits was from northwest (Bertran *et al.*, 2016; Borderie *et al.*, 2017). The Beauce
416 area corresponds to the southern margin of the Paris basin loess deposits which
417 show a clear gradient from North to South (Bertran *et al.*, 2016). The gradient of
418 loess that we observe in the Beauce region from north to south may be due to this. In
419 addition, the northern part is characterized by large flat plateaus where no erosion
420 occurred, except along the main deep valleys, whereas the southwestern part is
421 characterized by a local relief that may have induced erosion and redistribution
422 processes (Macaire, 1971). All these observations were confirmed by the expert
423 knowledge of the soil surveyors who did some traditional reconnaissance soil
424 mapping in this region. Interestingly, when doing reconnaissance maps at 1:250,000
425 the soil surveyors delineated small natural regions in order to create the broadest
426 geographical ensembles of the legend (Richer-de-Forges, 2008; Richer-de-Forges *et*

427 *al.*, 2008). Figure 9 shows these small natural regions drawn by the soil surveyors on
428 the study area. The comparison between Figure 8 and Figure 9 clearly shows that
429 the map of the TOL makes sense both from soil and physiography point of views.
430 One should keep in mind that 90% PI is a very large PI. Therefore, it is normal that
431 such wide ranges are found. Another reason for the large PI comes from the fact that
432 our map does not have very high model performance, and there is still a large room
433 to improve it. Useful outputs for irrigation or drainage management, however, are
434 maps of probability of exceeding a given depth for TOL in the study area (see an
435 example in the next section).

436 **3.5. Example of application**

437 One example of application is to map the probability of the TOL to exceed a
438 given depth or, on the contrary, to map the probability of the TOL to be less than a
439 given value. Figure 10 displays an example of these practical applications, which
440 extracts the probability of exceeding of 30 cm from the function between the TOL and
441 prediction quantile (from 0 to 100% with an interval of 2%) within the QRF model. The
442 soils that have a very low probability to exceed a 30 cm TOL are unsuitable for
443 conventional tillage and have a very low AWC. Therefore, optimizing the irrigation on
444 these soils should greatly save water.

445

446 **4. Conclusion**

447 In this study, we utilized a large soil dataset (10978 sampling sites) and 117
448 environmental covariates relevant to soil, organisms, relief, parent material and
449 spatial position for mapping thickness of loess at a regional scale. The 50 repeated
450 Quantile Random Forest had an average R^2 of 0.33, which was slightly better than
451 those obtained in most previous studies at regional or national scale (R^2 of
452 0.11~0.41). A PICP of 86.70% showed that around 86.70% of the validation samples
453 fall within the predefined 90% PIs, which indicated that the prediction uncertainty
454 produced by Quantile Random Forest was reasonable and can be properly used in
455 decision making of land management. The relative importance of environmental
456 covariates indicated the importance of elevation and gamma radiometry in modelling
457 thickness of loess and also proved the necessity of incorporating neighbour
458 information in relief and organisms for spatial modelling. The produced map of

459 thickness of loess and its 90% prediction intervals made sense from a soil science
460 perspective. This map can be further used for efficient irrigation management as well
461 as crop growth and yield modelling.

462 **Acknowledgement**

463 Dominique Arrouays is coordinator, Vera Leatitia Mulder is member, and Anne
464 C. Richer-de-Forges, and Guillaume Martelet are collaborators of the Research
465 Consortium GLADSOILMAP supported by “LE STUDIUM” Loire Valley Institute for
466 Advances Research Studies. This work was funded by the French National Research
467 Agency (Soilserv program, ANR 16 CE32 0005 01). Songchao Chen has received
468 the support of China Scholarship Council for 3 years’ Ph.D. study in INRAE and
469 Agrocampus Ouest (under grant agreement no. 201606320211). We also
470 acknowledge Dr. H el ene Tissoux (BRGM, France) for sharing her knowledge on
471 loess deposits in France and the Beauce region.

472

References

- Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., ..., Mendonca-Santos, M.d.L., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. In *Advances in agronomy* (Vol. 125, pp. 93-134). Academic Press.
- Behrens, T., MacMillan, R. A., Viscarra Rossel, R.A., Schmidt, K., Lee, J., 2019. Teleconnections in spatial modelling. *Geoderma*, 354, 113854.
- Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018a. Multi-scale digital soil mapping with deep learning. *Scientific reports*, 8(1), 1-9.
- Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018b. Multiscale contextual spatial modelling with the Gaussian scale space. *Geoderma*, 310, 128-137.
- Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155(3-4), 175-185.
- Bertran, P., Liard, M., Sitzia, L., Tissoux, H., 2016. A map of Pleistocene aeolian deposits in Western Europe, with special emphasis on France. *Journal of Quaternary Science*, 31(8), 844-856.
- Biswas, A., Cresswell, H.P., Chau, H.W., Viscarra Rossel, R.A., Si, B.C., 2013a. Separating scale-specific soil spatial variability: A comparison of multi-resolution analysis and empirical mode decomposition. *Geoderma*, 209, 57-64.
- Biswas, A., Cresswell, H.P., Viscarra Rossel, R.A., Si, B.C., 2013b. Characterizing scale - and location - specific variation in non - linear soil systems using the wavelet transform. *European Journal of Soil Science*, 64(5), 706-715.
- Borderie, Q., Chamaux, G., Roussaffa, H., Douard, M., Fencke, E., Rodot, M.-A., Perrichon, P., Selles, H., 2017. La couverture loessique d'Eure-et-Loir (France): Potentiel pédo-sédimentaire et organisation spatiale. *Quaternaire*, 28(3), 389-400.

- Bourennane, H., King, D., Chery, P., Bruand, A., 1996. Improving the kriging of a soil variable using slope gradient as external drift. *European Journal of Soil Science*, 47(4), 473-483.
- Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5-32.
- British Standards Institution, 1988. *BSI Standards Catalogue: 1988*. The Institution.
- Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62(3), 394-407.
- Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., ..., van Wesemael, B., 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147, 267-282.
- Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D., 2018. Fine resolution map of top-and subsoil carbon sequestration potential in France. *Science of the Total Environment*, 630, 389-400.
- Chen, S., Mulder, V.L., Martin, M.P., Walter, C., Lacoste, M., Richer-de-Forges, A.C. C., Arrouays, D., 2019. Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma*, 344, 184-194.
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, 8, 1991-2007.
- Copernicus, 2016. Contains Modified Copernicus Service Information.
<https://land.copernicus.eu/global/products/ndvi>.
- Doorenbos, J., Kassam, A.H., Bentvelder, C., Uittenboogaard, G., 1978. Yield response to water. *FAO Irrigation and Drainage Paper 33*. FAO, Rome, Italy, p. 144.
- Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment*, 218, 89-103.

- Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143(1-2), 180-190.
- Grundy, M.J., Viscarra Rossel, R.A., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L. J., 2015. Soil and landscape grid of Australia. *Soil Research*, 53(8), 835-844.
- Guerrero, E., Pérez, A., Arroyo, C., Equihua, J., Guevara, M., 2014. Building a national framework for pedometric mapping: soil depth as an example for Mexico. In: *GlobalSoilMap: Basis of the global spatial soil information system*, Taylor & Francis, CRC Press, London 103-108.
- Hagolle, O., Huc, M., Villa Pascual, D., Dedieu, G., 2015. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VEN μ S and Sentinel-2 images. *Remote Sensing*, 7(3), 2668-2691.
- Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Guevara, M.A., 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2), e0169748.
- Huang, J., Wu, C., Minasny, B., Roudier, P., McBratney, A.B., 2017. Unravelling scale-and location-specific variations in soil properties using the 2-dimensional empirical mode decomposition. *Geoderma*, 307, 139-149.
- IGN, 2011. BD ALTI[®] Version 2.0 - Descriptif de Contenu. Available online at: http://professionnels.ign.fr/sites/default/files/DC_BDALTI_2-0.pdf
- Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1), 95.
- IUSS Working Group, WRB, 2006. World reference base for soil resources. *World Soil Resources Report*, 103.
- Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A., 2015. Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. *Soil Research*, 53(8), 932-955.

- King, D., Daroussin, J., Tavernier, R., 1994. Development of a soil geographic database from the Soil Map of the European Communities. *CATENA*, 21(1), 37-56.
- Kuhn, M., 2020. caret: Classification and Regression Training. R package version 6.0-85. <https://CRAN.R-project.org/package=caret>
- Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. Evaluating large-extent spatial modeling approaches: A case study for soil depth for France. *Geoderma Regional*, 7(2), 137-152.
- Leenaars, J.G., Claessens, L., Heuvelink, G.B., Hengl, T., González, M.R., van Bussel, L.G., ..., Cassman, K.G., 2018. Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. *Geoderma*, 324, 18-36.
- Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., Viscarra Rossel, R.A., 2019. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma*, 335, 47-56.
- Loiseau, T., Chen, S., Mulder, V. L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., ..., Arrouays, D., 2019. Satellite data integration for soil clay content modelling at a national scale. *International Journal of Applied Earth Observation and Geoinformation*, 82, 101905.
- Lombardo, L., Saia, S., Schillaci, C., Mai, P.M., Huser, R., 2018. Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. *Geoderma*, 318, 148-159.
- Lorain, J.M., 1973. La géologie du calcaire de Beauce. *Bulletin de Liaison des Laboratoires des Ponts et Chaussées. Special Issue*, 78.
- Macaire, J.M., 1971. Etude sédimentologique des formations superficielles sur le tracé de l'autoroute A10 entre Artenay et Meung sur Loire (PhD thesis). University of Orleans, France.
- Martelet, G., Drufin, S., Tourlière, B., Saby, N. P. A., Perrin, J., DeParis, J., Prognon, J. F., Jolivet, C., Ratié, C., Arrouays, D., 2013. Regional regolith parameters

- prediction using the proxy of airborne gamma ray spectrometry. *Vadose Zone Journal*, 12(4), 1-14.
- Martelet, G., Nehlig, P., Arrouays, D., Messner, F., Tourlière, B., Laroche, B., Deparis, J., Saby, N., Richer-de-Forges, A.C., Jolivet, C., Ratié, C., 2014. Airborne gamma-ray spectrometry: potential for regolith-soil mapping and characterization. In: *GlobalSoilMap: Basis of the global spatial soil information system*. Taylor & Francis, CRC Press, London 401-408.
- McBratney, A.B., Santos, M.d.L., Minasny, B., 2003. On digital soil mapping. *Geoderma*, 117(1-2), 3-52.
- Meinshausen, N., 2006. Quantile regression forests. *Journal of Machine Learning Research*, 7, 983-999.
- Meinshausen, N., 2017. *quantregForest: Quantile Regression Forests*. R package version 1.3-7.
- Ménillet, F., 1974. Etude pétrographique et sédimentologique des calcaires d'Etampes et de Beauce. Formations dulcaquicoles du Stampien supérieur à l'Aquitainien dans le Bassin de Paris. Thesis, University of Paris-Sud (Orsay).
- Møller, A.B., Beucher, A.M., Pouladi, N., Greve, M.H., 2019. Oblique geographic coordinates as covariates for digital soil mapping. *SOIL Discussions*, 1-20.
- Mulder, V. L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016. *GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth*. *Science of the total environment*, 573, 1352-1369.
- Nicoullaud, B., Darhout, R., Duval, O., 1995. Etude de l'enracinement du blé tendre d'hiver et du maïs dans les sols argilo-limoneux de Petite Beauce. *Etude et Gestion des Sols*, 2, 183-200.
- Ould-Mohamed, S., Bruand, A., Bruckler, L., Bertuzzi, P., Guillet, B., Raison, L., 1997. Estimating long-term drainage at a regional scale using a deterministic model. *Soil Science Society of America Journal*, 61, 1473-1482.
- Paroissien, J. B., Saby, N., de Forges, A. R., Arrouays, D., Louis, B., 2014. Populating soil maps with legacy data from a soil testing databases. In:

- GlobalSoilMap: Basis of the global spatial soil information system. Taylor & Francis, CRC Press, London 319-324.
- R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Richer-de-Forges, A.C., 2008. Notice explicative de la carte des pédopaysages du Loiret à 1/250 000 (Référentiel Régional Pédologique de la région Centre). INRA InfoSol. (in French)
- Richer-de-Forges, A.C., Desbourdes, S., Lehmann, S., 2008. Référentiel Régional Pédologique de la région Centre : carte des pédopaysages du Loiret à 1/250 000. INRA InfoSol. (in French)
- Shangguan, W., Hengl, T., de Jesus, J. M., Yuan, H., Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems*, 9(1), 65-88.
- Tetegan, M., Richer-de-Forges, A.C., Verbeque, B., Nicoullaud, B., Desbourdes, C., Bouthier, A., Arrouays, D., Cousin, I., 2015. The effect of soil stoniness on the estimation of water retention. *Catena*, 129, 96-102.
- Vaudour, E., Gomez, C., Fouad, Y., Lagacherie, P., 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. *Remote Sensing of Environment*, 223, 21-33.
- Vaysse, K., Lagacherie, P., 2015. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). *Geoderma Regional*, 4, 20-30.
- Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, 291, 55-64.
- Walvoort, D., Brus, D., de Gruijter, J., 2020. Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata. R package version 0.3-9. <https://CRAN.R-project.org/package=spcosa>
- Zhang, W., Hu, G., Sheng, J., Weindorf, D.C., Wu, H., Xuan, J., ..., Gu, Z., 2018. Estimating effective soil depth at regional scales: Legacy maps versus

environmental covariates. *Journal of Plant Nutrition and Soil Science*, 181(2), 167-176.

Zhao, R., Biswas, A., Zhou, Y., Zhou, Y., Shi, Z., Li, H., 2018. Identifying localized and scale-specific multivariate controls of soil organic matter variations using multiple wavelet coherence. *Science of The Total Environment*, 643, 548-558.

Zhou, Y., Biswas, A., Ma, Z., Lu, Y., Chen, Q., Shi, Z., 2016. Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain. *Geoderma*, 271, 71-79.

Figures

Figure 1 Study area and soil sampling sites, the Beauce area that locates at the middle Loire catchment, central France

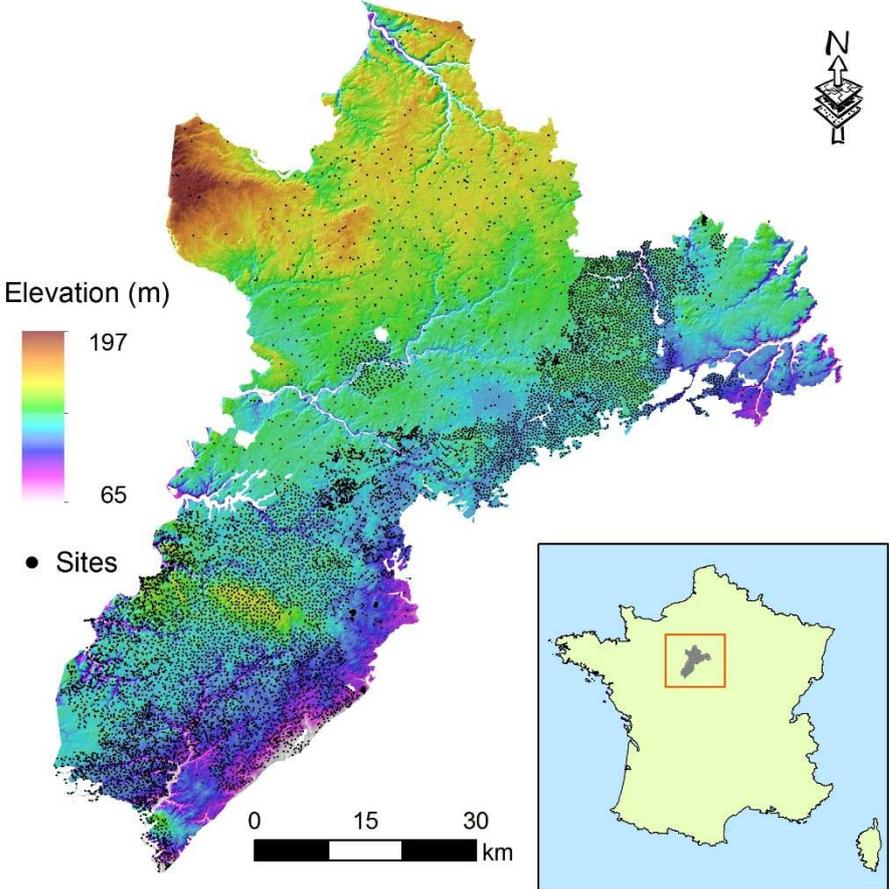


Figure 2 Examples of environmental covariates

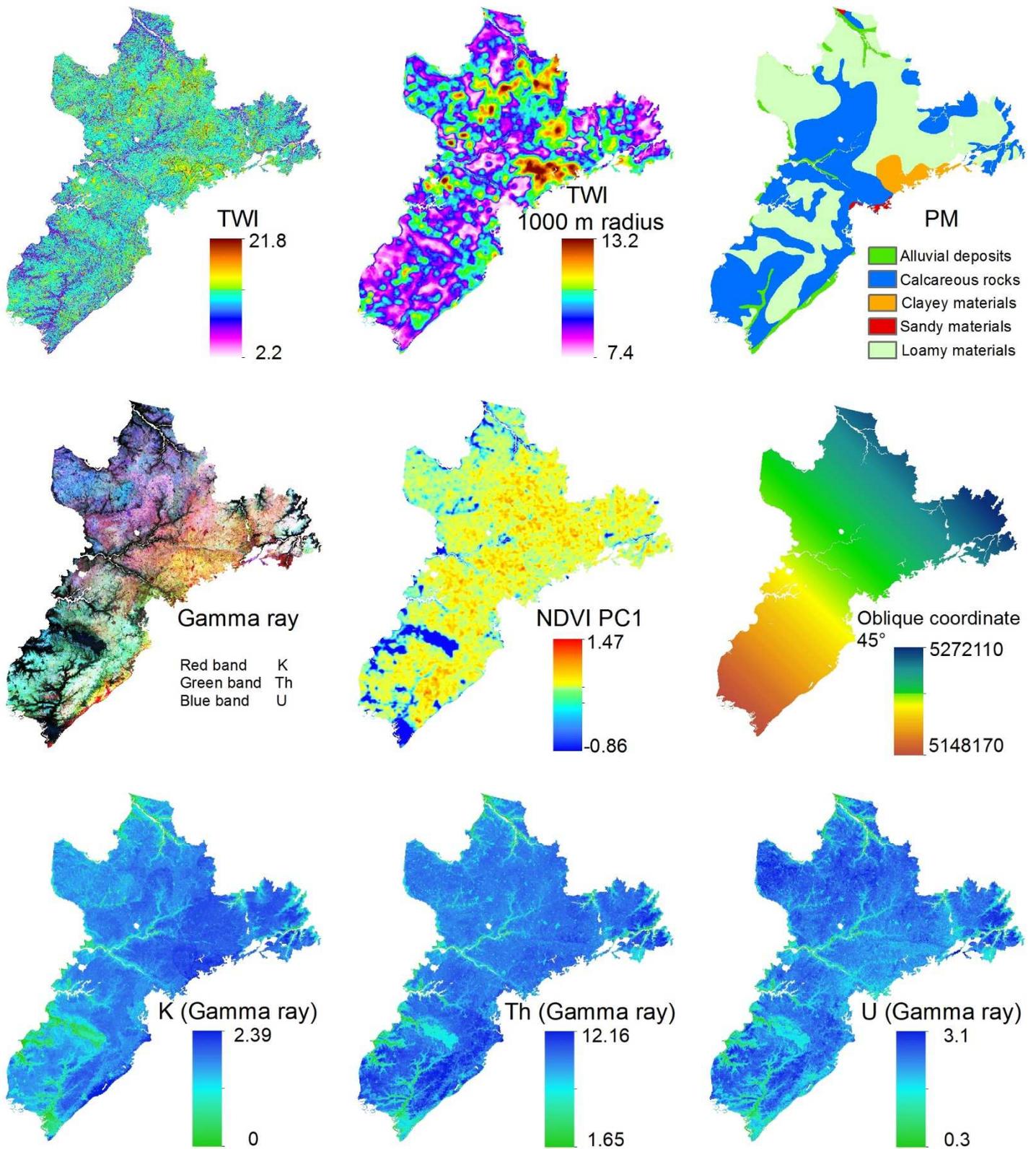


Figure 3 Compact equal area geographical strata

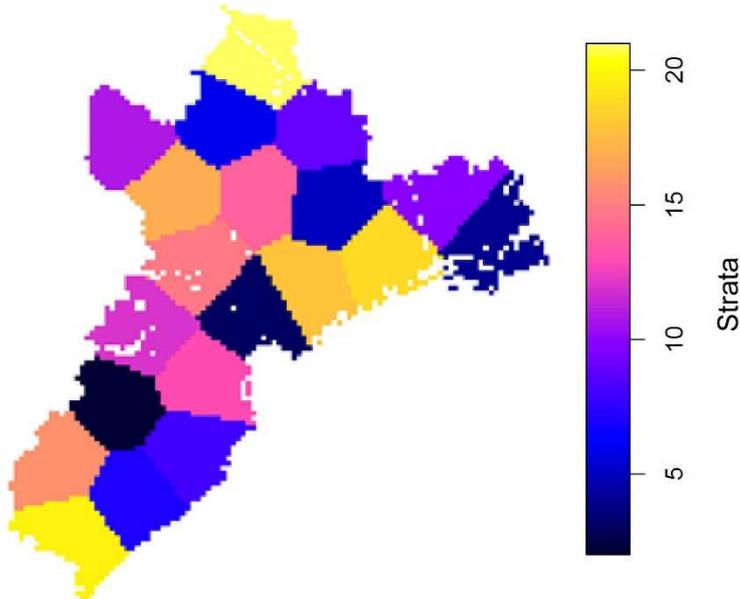


Figure 4 Boxplot of TOL for each compact equal area geographical strata. The number of sampling sites is indicated in blue.

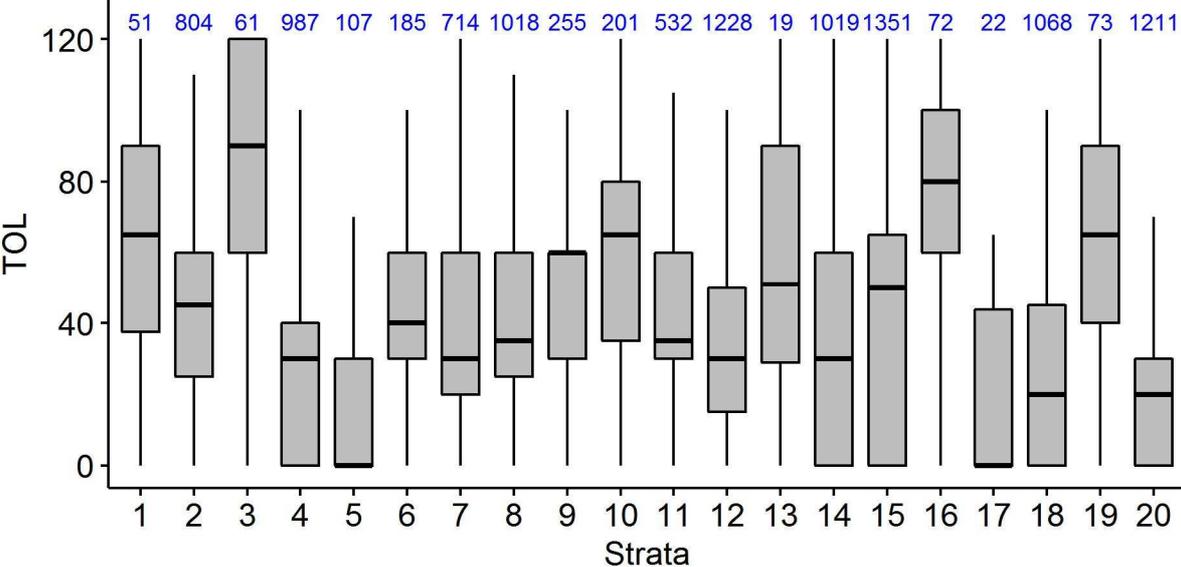


Figure 5 Pearson correlation coefficient (r) between TOL and top correlated environmental covariates ($r > 0.15$ or $r < -0.15$).

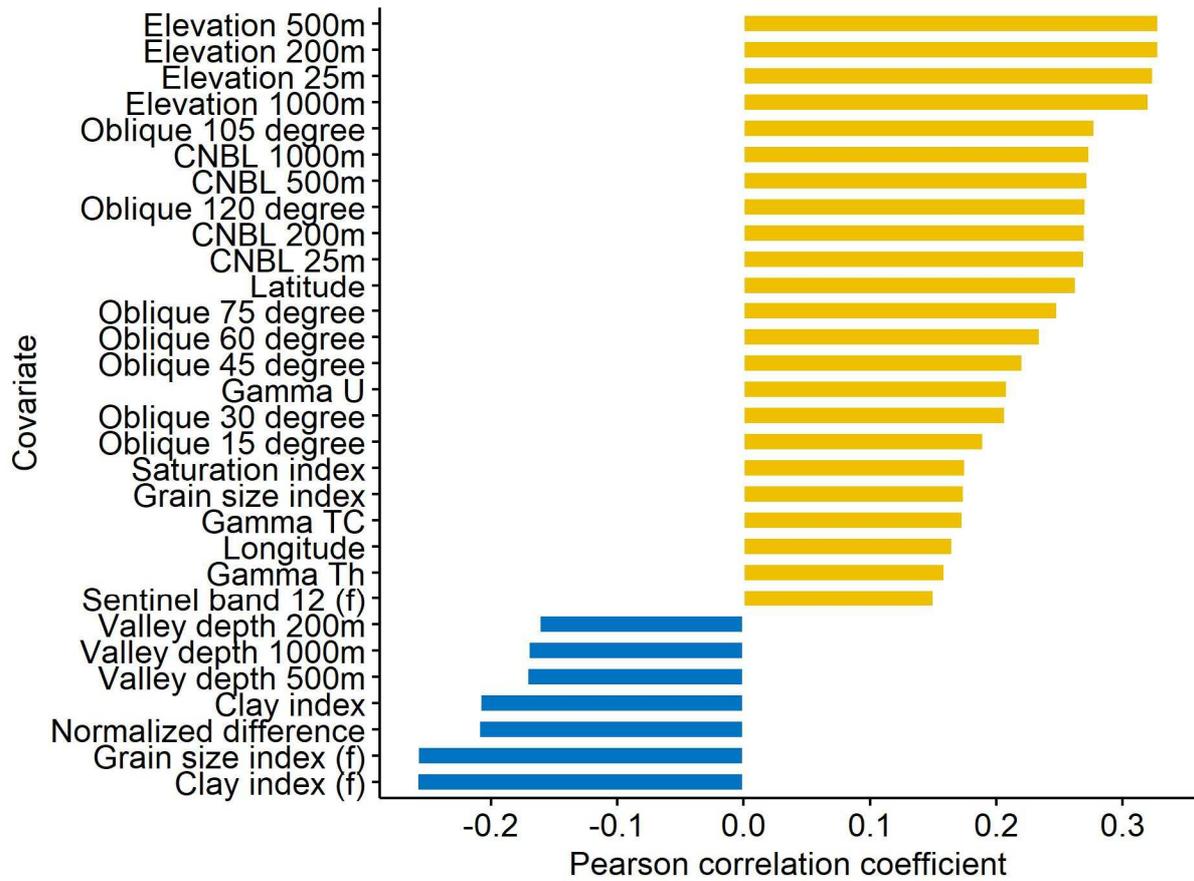


Figure 6 Model performance of 50 repeats evaluated by R^2 (a), RMSE (b), bias (c) and coverage of PICP (d) on validation set

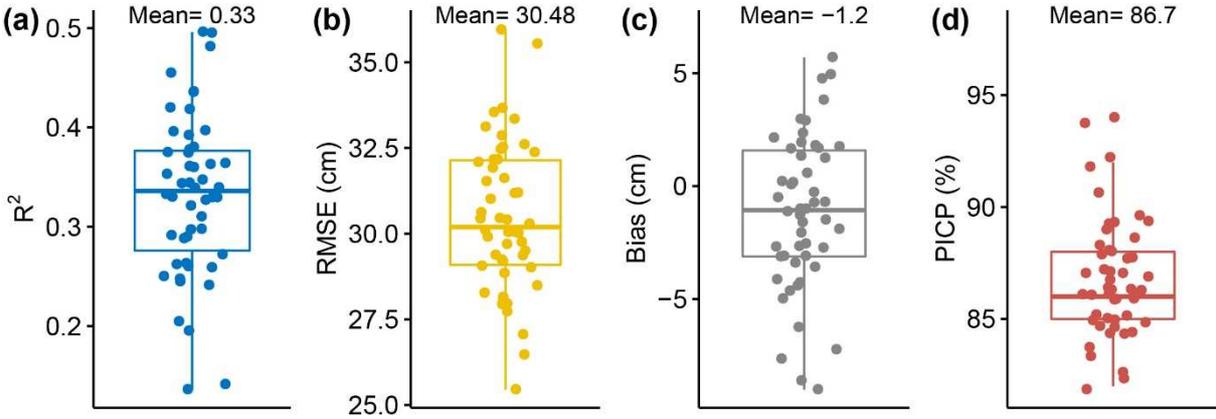


Figure 7 Relative importance of environmental covariates in Quantile Random Forest (average of 50 repeats). Only the top 30 variables are shown here.

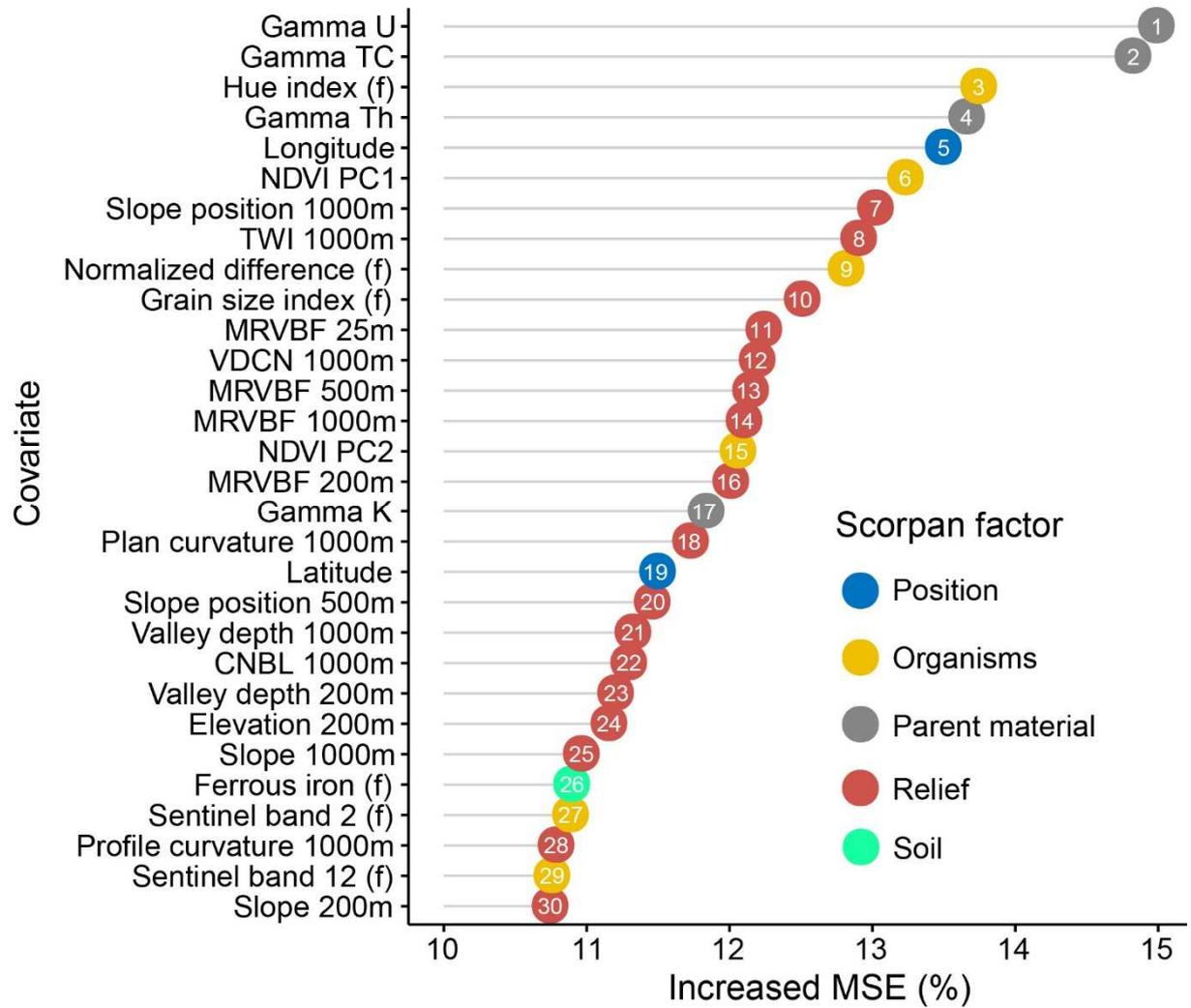


Figure 8 Spatial distribution of the thickness of loess and its 90% prediction intervals

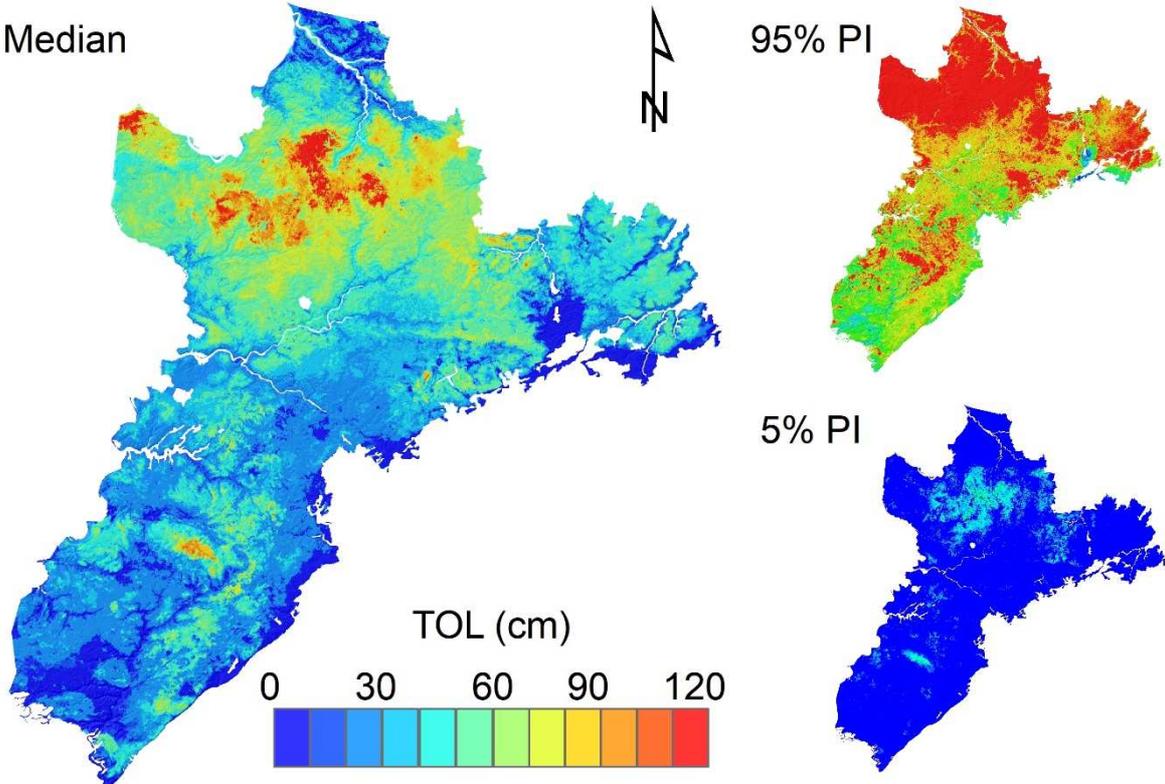


Figure 9 Natural regions delineated by soil surveyors (Richer-de-Forges, 2008; Richer-de-Forges *et al.*, 2008)

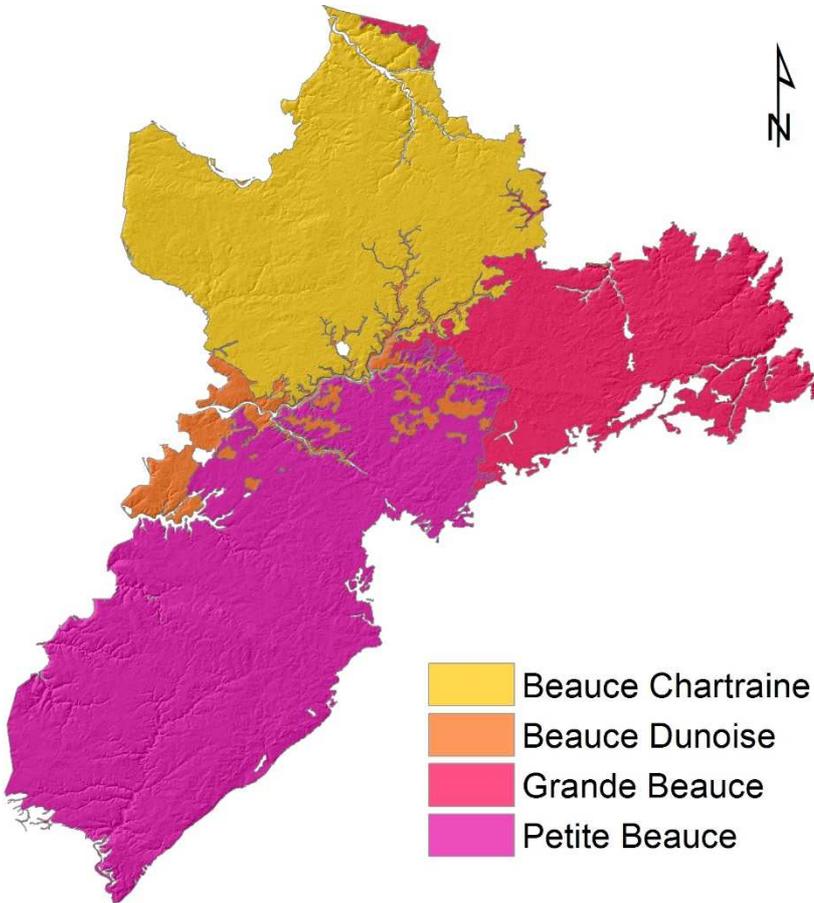
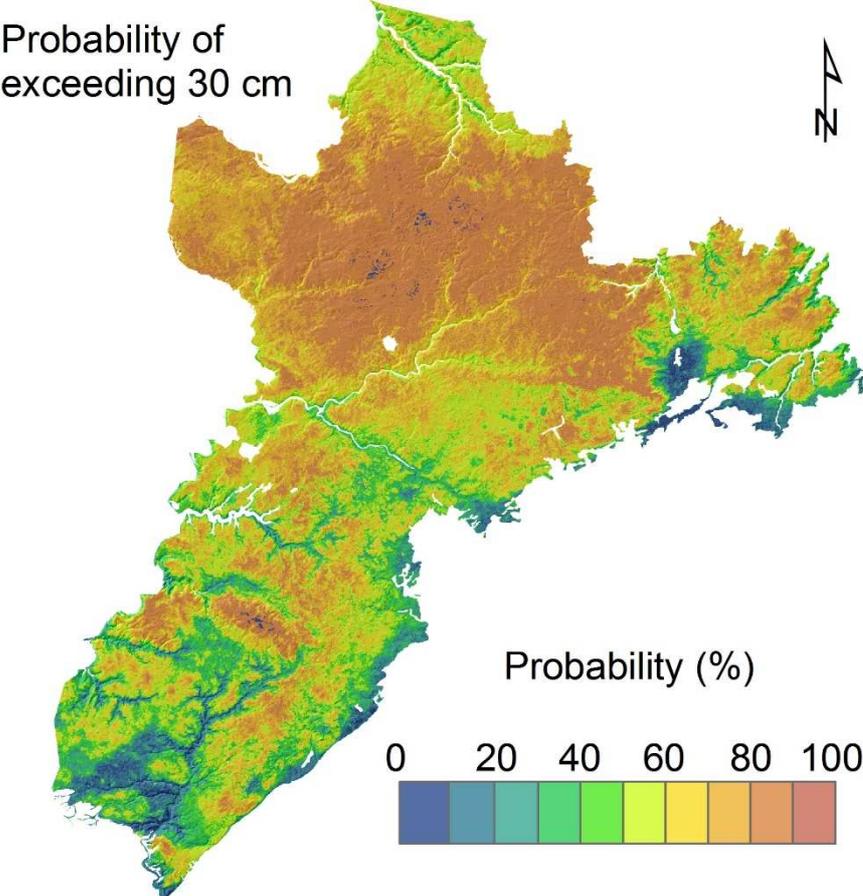


Figure 10 The probability of exceeding 30 cm for TOL in the study area. The probability at 30 cm is extracted from the probability distribution of Quantile Random Forest for each pixel.



Tables

Table 1 Environmental covariates used for digital soil mapping

Variable	Number	Resolution	Scorpan factor	Reference
Channel network base level	4	25 m	Relief ^a	IGN, 2011
Elevation	4	25 m	Relief	IGN, 2011
Multiresolution index of valley bottom flatness	4	25 m	Relief	IGN, 2011
Plan curvature	4	25 m	Relief	IGN, 2011
Profile curvature	4	25 m	Relief	IGN, 2011
Slope	4	25 m	Relief	IGN, 2011
Slope position	4	25 m	Relief	IGN, 2011
Slope length	4	25 m	Relief	IGN, 2011
Terrain wetness index	4	25 m	Relief	IGN, 2011
Valley depth	4	25 m	Relief	IGN, 2011
Vertical distance to channel network	4	25 m	Relief	IGN, 2011
Soil type	1	1:1000000	Soil	King <i>et al.</i> (1995)
Parent material	1	1:1000000	Parent material	King <i>et al.</i> (1995)
Gamma radiometric (K, U, Th, TC)	4	200 m	Parent material	Martelet <i>et al.</i> (2014)
Land cover and probability	10	10, 20, 60 m	Organisms	Inglada <i>et al.</i> (2017)
Sentinel 2 spectral bands and indices	42	90 m	Organisms	Loiseau <i>et al.</i> (2019)
First three PCs of monthly NDVI ^b	3	300, 500 m	Organisms	Loiseau <i>et al.</i> (2019)
Coordinates (Latitude, Longitude)	2	25 m	Position	IGN, 2011
Oblique coordinates ^c	10	25 m	Position	Møller <i>et al.</i> (2019)

^a For all the covariates in relief factor, except for the original products, their local mean values with radius at 200, 500 and 1000 m are also calculated by convolution circular windows. ^b PCs, principal components; NDVI, normalized difference vegetation index. ^c Oblique coordinates at angles of 15°, 30°, 45°, 60°, 75°, 105°, 120°, 135°, 150° and 165° are produced

Table 2 Random Forest model performance over the predictor subset size using recursive feature selection

Subset size	RMSE	R ²	Selected
5	28.70	0.1808	No
10	27.07	0.2712	No
15	26.72	0.2904	No
20	26.52	0.3007	No
40	26.50	0.3022	No
60	26.50	0.3019	No
80	26.44	0.3052	Yes
100	26.50	0.3025	No
117	26.48	0.3038	No

Table 3 Statistics of the thickness of loess (in cm)

Variable	Number	Minimum	Q1	Median	Mean	Q3	Maximum	Skewness	Kurtosis
TOL	10978	0	0	30	36.46	60	120	0.66	2.85

Q1, the first quartile; Q3, the third quartile.

Table 4 A summary of the model performance of soil thickness (or soil depth) mapping from regional to global scales

Reference	Location	R ²
<i>Regional scale</i>		
Kidd et al. (2015)	Tasmania, Australia	0.16
Vaysse and Lagacherie (2015)	Languedoc-Roussillon, France	0.23
Zhang et al. (2018)	Xinjiang, China	0.28
This study	Beauce, France	0.34
<i>National scale</i>		
Guerrero et al. (2014)	Mexico	0.41
Lacoste et al. (2016)	France	0.22
Mulder et al. (2016)	France	0.11
<i>Global scale</i>		
Hengl et al. (2017)	Globe	0.57
Shangguan et al. (2017)	Globe	0.59