



HAL
open science

Probability mapping of soil thickness by random survival forest at a national scale

Songchao Chen, Vera Leatitia Mulder, Manuel M. Martin, Christian Walter, Marine Lacoste, Anne C Richer-De-Forges, Nicolas P.A. Saby, Thomas Loiseau, Bifeng Hu, Dominique Arrouays

► To cite this version:

Songchao Chen, Vera Leatitia Mulder, Manuel M. Martin, Christian Walter, Marine Lacoste, et al.. Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma*, 2019, 344, pp.184-194. 10.1016/j.geoderma.2019.03.016 . hal-02111857

HAL Id: hal-02111857

<https://hal.science/hal-02111857v1>

Submitted on 22 Oct 2021

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 - NonCommercial 4.0 International License

1 **Title:** Probability mapping of soil thickness by random survival forest at a national
2 scale

3

4 **Authors:**

5 Songchao Chen ^{a, b}. songchao.chen@inra.fr

6 Vera Leatitia Mulder ^c. titia.mulder@wur.nl

7 Manuel P. Martin ^a. manuel.martin@inra.fr

8 Christian Walter ^b. christian.walter@agrocampus-ouest.fr

9 Marine Lacoste ^d. marine.lacoste@inra.fr

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

11 Nicolas P.A. Saby ^a. nicolas.saby@inra.fr

12 Thomas Loiseau ^a. thomas.loiseau@inra.fr

13 Bifeng Hu ^{a, d, e}. bifeng.hu@inra.fr

14 Dominique Arrouays ^a. dominique.arrouays@inra.fr

15 **Affiliations:**

16 ^a INRA, Unité InfoSol, 45075 Orléans, France

17 ^b SAS, INRA, Agrocampus Ouest, 35042 Rennes, France

18 ^c Soil Geography and Landscape group, Wageningen University, PO Box 47 6700

19 AA Wageningen, The Netherlands

20 ^d INRA, Unité Science du sol, 45075 Orléans, France

21 ^e Sciences de la Terre et de l'Univers, Orléans University, 45067 Orléans, France

22 **Corresponding author:**

23 Songchao Chen: songchao.chen@inra.fr

24 Postal address: INRA, Unité InfoSol, 2163 Avenue de la Pomme de Pin, CS 40001

25 Ardon, 45075 Orléans, France

26 Telephone: +33(0)602142667

27 **Abstract:**

28 Soil thickness (ST) is a crucial factor in earth surface modelling and soil storage
29 capacity calculations (e.g., available water capacity and carbon stocks). However, the
30 observed depths recorded in soil information systems for some profiles are often less
31 than the actual ST (i.e., right censored data). The use of such data will negatively
32 affect model and map accuracy, yet few studies have been done to resolve this issue
33 or propose methods to correct for right censored data. Therefore, this work
34 demonstrates how right censored data can be accounted for in the ST modelling of
35 mainland France. We propose the use of Random Survival Forest (RSF) for ST
36 probability mapping within a Digital Soil Mapping framework. Among 2109 sites of the
37 French Soil Monitoring Network, 1089 observed STs were defined as being right
38 censored. Using RSF, the probability of exceeding a given depth was modelled using
39 freely available spatial data representing the main soil-forming factors. Subsequently,
40 the models were extrapolated to the full spatial extent of mainland France. As
41 examples, we produced maps showing the probability of exceeding the thickness of
42 each *GlobalSoilMap* standard depth: 5, 15, 30, 60, 100, and 200 cm. In addition, a
43 bootstrapping approach was used to assess the 90% confidence intervals. Our
44 results showed that RSF was able to correct for right censored data entries occurring
45 within a given dataset. RSF was more reliable for thin (0.3 m) and thick soils (1 to 2
46 m), as they performed better (overall accuracy from 0.793 to 0.989) than soils with a
47 thickness between 0.3 and 1 m. This study provides a new approach for modelling
48 right censored soil information. Moreover, RSF can produce probability maps at any
49 depth less than the maximum depth of the calibration data, which is of great value for
50 designing additional sampling campaigns and decision making in geotechnical
51 engineering.

- 52 **Keywords:** Soil thickness modelling; Right censored data; Random Survival Forest;
- 53 GlobalSoilMap; Probability mapping.

54 **1. Introduction**

55 Soils are of great importance in supporting, provisioning, and regulating
56 ecosystem services, such as food production and climate change mitigation (Clothier
57 et al., 2011; Keesstra et al., 2016; Millennium Ecosystems Assessment, 2005). As
58 stated by recent studies (e.g., Bouma, 2018; Groshans et al., 2018; Marx et al., 2019),
59 there is a rising demand for up-to-date and ecosystem service relevant soil
60 information. Therefore, substantial effort is needed to communicate soil information
61 among diverse audiences and produce fine resolution soil maps to support practical
62 land management. In this study, we use the *GlobalSoilMap* project specifications.
63 These specifications focus on delivering consistently produced high-resolution soil
64 property information throughout the world by predicting mean values and their
65 prediction intervals (PIs) (Arrouays et al., 2014a; Arrouays et al., 2014b; Sanchez et
66 al., 2009). Among the twelve soil properties to be predicted following the
67 recommendations of *GlobalSoilMap*, soil thickness (ST) is a key property. In this
68 study, in line with *GlobalSoilMap*, ST is defined as ‘the depth (cm) from the soil
69 surface to the lithic or a paralithic contact’ (Soil Survey Division Staff, 1993). The ST
70 is highly relevant for soil hydro-mechanical modelling (Tesfa et al., 2009; Wang et al.,
71 2006), soil erosion impact, landscape evolution, vegetation growth (Heimsath et al.,
72 2001; Meyer et al., 2007), and for calculating soil functions (e.g., available water
73 capacity (Leenaars et al., 2018; Román Dobarco et al., 2019), soil structure (Rabot et
74 al., 2018; Vogel et al., 2018), and soil organic carbon stocks (Batjes, 1996; Chen et
75 al., 2019)). Despite the great importance of accurate ST information, the large spatial
76 variability and high cost of ST measurements make ST determination difficult
77 (Lacoste et al., 2016). Discordance in the definition of ST also hampers ST modelling,
78 especially when data are collected from various projects (Lacoste et al., 2016). The

79 observed ST recorded in soil information systems for some profiles are often less
80 than the actual ST (i.e., right censored data).

81 ST results from the mass balance between soil formation from the bedrock
82 and soil transport by erosion and sedimentation (Heimsath et al., 1997; Heimsath et
83 al., 1999); thus, it varies as a function of physical, chemical, and biological processes
84 (Román et al., 2018). ST can be related to these processes by modeling the
85 relationship between the main soil-forming factors, i.e., Jenny's Soil-Landscape
86 paradigm (Jenny, 1941): parent material, climatic conditions, organisms, terrain relief,
87 and time (Dietrich et al., 1995; Minasny and McBratney, 1999). More recently,
88 McBratney et al. (2003) formulated the concept of the SCORPAN model, which also
89 includes also soil information and spatial location.

90 Various approaches for ST modelling and mapping have relied on modelling
91 the relationship between the main soil forming factors. The majority of these
92 approaches can be broadly classified into two groups: 1) physically based and
93 mechanistic models, which predict ST using soil process models based on the rates
94 of weathering, denudation, and accumulation (Bonfatti et al., 2018; Dietrich et al.,
95 1995; Minasny and McBratney, 1999; Pelletier and Rasmussen, 2009); and 2)
96 empirical models, including statistical and geostatistical methods (Kuriakose et al.,
97 2009). These models rely on the empirical relationships between ST and explanatory
98 covariates of inferential attributes (e.g., plant species, precipitation, and parent
99 material).

100 For the latter, a wide range of statistical methods have been previously applied
101 in ST modelling, including canonical correspondence analysis and principal
102 component analysis (Odeh et al., 1991), multiple linear regression (Moore et al.,
103 1993), expert knowledge and fuzzy logic (Zhu et al., 2001), Generalized Additive

104 Models and Random Forest (Tesfa et al., 2009), and Cubist and Gradient Boosting
105 Modelling (Lacoste et al., 2016; Mulder et al., 2016a).

106 Within the field of geostatistics, various kriging techniques have often been
107 used to predict and spatially interpolate ST from point samples. Ordinary Kriging was
108 most commonly used among these kriging techniques (Penížek and Borůvka, 2006;
109 Vanwalleggem et al., 2010). The prediction variance was typically reduced when
110 including additional prediction variables using regression kriging (Kuriakose et al.,
111 2009; Odeh et al., 1995) or Kriging with External Drift (Bourennane et al., 1996;
112 Kempen et al., 2015).

113 None of the studies referred to above addressed the issue of having right
114 censored data entries in their soil databases. However, it is often the case that the
115 actual ST is thicker than the observed ST, which can mainly be attributed to practical
116 constraints, such as the standard auger length (120 cm), and time constraints. In fact,
117 in soil sciences very few studies consider the effect of right censored data; the issue
118 is often ignored or processed by adding a fixed value (e.g., 30 cm) in ST modelling
119 (Knotters et al., 1995; Vaysse and Lagacherie, 2015; Lacoste et al., 2016;
120 Shangguan et al., 2017). Some previous works dealt with left censored data,
121 especially regarding data below detection limits (e.g., de Oliveira, 2005; Fridley and
122 Dixon, 2007; Orton et al., 2009; Orton et al., 2012; Villaneau et al., 2011). Ignoring
123 the presence of right censored data entries within a database and relying on the
124 observed ST for those entries will result in an underestimation of modelled ST
125 (Vaysse and Lagacherie, 2015; Shangguan et al., 2017).

126 However, right censored data are commonly used in statistics and medical
127 research, especially in survival analysis. Several models have been used to deal with
128 right censored data in survival analysis, including the Kaplan Meier method (Kaplan

129 and Meier, 1958), Cox regression (Andersen and Gill, 1982), and Random Survival
130 Forest (RSF, Ishwaran et al., 2008). The Kaplan Meier method and Cox regression
131 mainly deal with linear effects, but RSF is capable of handling complex non-linear
132 effects that may exist between predictor variables (Mogensen et al., 2012). Therefore,
133 as previously suggested by Styc and Lagacherie (2016), RSF may have the best
134 potential for identifying and correcting right censored data used for Digital Soil
135 Mapping (DSM).

136 In this study, the potential of RSF was evaluated for ST mapping in mainland
137 France. The main objectives of this study are noted below:

- 138 1) Apply RSF for mapping the probability of exceeding a certain ST using both
139 actual and right censored ST data from the French Soil Monitoring Network
140 (RMQS) and
- 141 2) Derive the 90% confidence intervals of the specific ST using bootstrapping.

142

143 **2. Material and methods**

144 2.1. Soil dataset

145 We used ST data from the RMQS soil database that were gathered between
146 2001 and 2009 (Jolivet et al., 2006), covering different soil, climate, relief, and land
147 cover conditions (Fig. 1). The RMQS dataset is based on a 16 km × 16 km square
148 grid where all sites are selected at the centre of each grid cell. When sampling the
149 exact location was not possible, a site was selected as close as possible to the grid
150 centre. A soil pit was dug, and the surrounding information (land use and
151 geomorphology) and a detailed description of the soil profile were recorded for each
152 site, including soil horizon depth and ST. Auger boring was recommended (but not
153 mandatory) to complete the soil profile when the soil pit was not thick enough to

154 determine the ST. For more detailed information about the soil sampling design and
155 laboratory analysis, see Chen et al. (2018). Among 2109 RMQS sites, ST was
156 explicitly recorded for 1020 sites (down to a lithic or paralithic contact), while the
157 remaining 1089 sites were right censored data. The ST for nine RMQS sites was set
158 to 0, as these sites were identified as mountainous sites with bare rock.

159

160 2.2. Exhaustive covariates

161 We used a DSM framework (McBratney et al., 2003) to model the relationships
162 between ST and ancillary covariates (Table 1). These covariates cover a series of
163 soil formation related environmental factors, including soil, climate, organisms, relief
164 and parent material. Before modelling, these covariates were re-projected to Lambert
165 93 (official projection for mainland France) and resampled to a 90 m resolution (in
166 raster format) using a bilinear interpolation (numeric covariates) or nearest neighbour
167 (categorical covariates).

168

169 2.3. Random survival forest for probability modelling of soil thickness

170 2.3.1. General introduction

171 RSF is an ensemble tree method for modelling right censored survival data
172 (Ishwaran et al., 2008). RSF is an extension of Breiman's (2001) random forest (RF),
173 known as an ensemble learning approach that is improved by injecting randomization
174 into the base-learning process. For example, using a human analogy, a specific
175 status introduced in RSF aims to distinguish whether it is a death case (status = 1) or
176 a survival case (status = 0) at a given observed time. The RSF does this by
177 estimating the presence-absence probability. Applying this censoring concept to ST,
178 at a given site there are two possibilities for a thickness to be recorded: i) the actual

179 ST has been observed down to lithic or paralithic contact as defined before (status =
180 1) or ii) the lithic or paralithic contact has not been reached and the actual ST
181 remains unknown (status = 0). For the latter, it is only known that the actual ST is
182 greater than the observed ST. RSF uses a new type of predicted outcome that
183 contains a cumulative hazard function (CHF, see Section 2.3.2).

184

185 2.3.2. Random survival forest model fitting

186 RSF model fitting involves the following steps (Ishwaran et al., 2008), as
187 shown in Fig. 2.

- 188 1) Select *n*tree bootstrapped samples from the calibration data. Approximately 37%
189 (e^{-1}) of the calibration data are excluded in each bootstrapped sample, which are
190 so-called out-of-bag (OOB) data.
- 191 2) Grow a survival tree for each bootstrapped sample. At each node of the survival
192 tree, randomly select *m*try covariates for splitting the data. Survival splitting
193 criteria are then used, and each node is split on that covariate, which maximizes
194 survival differences across sub-nodes.
- 195 3) Grow the survival tree to full size under the constraint that a terminal node should
196 have no less than *nodesize* unique actual ST samples.
- 197 4) Calculate a CHF for each survival tree and obtain the ensemble CHF by
198 averaging all the survival trees for each sample.
- 199 5) Calculate the prediction error of the ensemble CHF based on OOB data.

200

201 *Ensemble cumulative hazard function and ensemble survival function*

202 Constructing the ensemble CHF is crucial for RSF. Hereafter, we provide
203 details about the procedure for a better understanding.

204 For a survival tree, let $(ST_{1,h}, \delta_{1,h}), \dots, (ST_{n(h),h}, \delta_{n(h),h})$ be the observed ST and
 205 the 0–1 censoring status (δ) for n samples in a terminal node h . Here, let $ST_{1,h} <$
 206 $ST_{2,h} < \dots < ST_{n(h),h}$ be the different observed ST in the terminal node. The CHF
 207 estimate for h is then defined by the Nelson–Aalen estimator \hat{H}_h :

$$208 \quad \hat{H}_h(st) = \sum_{st_{l,h} \leq st} \frac{a_{l,h}}{Y_{l,h}}, \quad (1)$$

209 where $a_{l,h}$ and $Y_{l,h}$ are the number of actual ST samples and all samples at observed
 210 ST $st_{l,h}$, respectively. All the samples within the terminal node h have the same CHF.

211 Each sample i has a m -dimensional covariate \mathbf{x}_i that will belong to a unique
 212 terminal node h . Therefore, the CHF for i is the Nelson–Aalen estimator for \mathbf{x}_i '
 213 terminal node:

$$214 \quad H(st|\mathbf{x}_i) = \hat{H}_h(st). \quad (2)$$

215 Equation 2 describes the CHF from an individual tree. The ensemble CHF is
 216 computed by averaging over $ntree$ trees. The bootstrap ensemble CHF for sample i is
 217 defined below (the definition of OOB ensemble CHF please refer to Ishwaran et al.,
 218 2008):

$$219 \quad H_e(st|\mathbf{x}_i) = \frac{1}{ntree} \sum_{n=1}^{ntree} H_n(st|\mathbf{x}_i), \quad (3)$$

220 where $H_n(st|\mathbf{x})$ is the CHF for a tree grown from the n^{th} bootstrap sample.

221 The survival function is a probability density function that describes the
 222 survival probability at a given ST. In RSF, the ensemble survival function (S_e) could
 223 be derived from ensemble CHF (Mogensen et al., 2012):

$$224 \quad S_e(st|\mathbf{x}_i) = \exp \left\{ -\frac{1}{ntree} \sum_{n=1}^{ntree} H_n(st|\mathbf{x}_i) \right\}. \quad (4)$$

225 Here, the survival probability at a given ST is equal to the probability of
 226 exceeding a given ST or censored probability at a given ST. The probability of

227 exceeding a given ST ranges from 0 to 1, and when it is close to 1, the location has a
 228 high probability of being censored. Therefore, in this latter case, the actual ST has a
 229 high probability of being thicker than the censored ST.

230

231 *Node splitting rule*

232 The node splitting rule is another important parameter in RSF. There are
 233 several choices for splitting rules, including the log-rank splitting rule, conservation
 234 splitting rule, log-rank score rule, and fast approximation to the log-rank splitting.
 235 Here, the log-rank splitting rule is used as the default splitting rule, as suggested by
 236 Ishwaran et al. (2008). We define $st_1 < st_2 < \dots < st_l$ as the ST intervals and $x_{i,j}$ and
 237 $a_{i,j}$ as the number of samples and number of actual ST samples at ST st_i in the sub-
 238 nodes j (1 or 2), respectively. Here, $x_i = x_{i,1} + x_{i,2}$ and $a_i = a_{i,1} + a_{i,2}$. The log-rank test
 239 for a split at the value n of the covariate c is defined as

$$240 \quad L(c, n) = \frac{\sum_{i=1}^l (a_{i,1} - x_{i,1} \frac{a_i}{x_i})}{\sqrt{\sum_{i=1}^l \frac{x_{i,1}}{x_i} (1 - \frac{x_{i,1}}{x_i}) (\frac{x_i - a_i}{x_i - 1}) a_i}}, \quad (5)$$

241 where the value $|L(c, n)|$ is the measure of node split, and $x_{i,1}$ and $a_{i,1}$ are the number
 242 of samples and number of actual ST samples, respectively, at ST st_i when c is less
 243 than n . The larger the $|L(c, n)|$ value, the larger the difference between two sub-
 244 nodes and a better split. The best split at each node is determined by searching the
 245 optimized covariate c^* and split value n^* to maximize the $|L(c, n)|$ value.

246

247 *Prediction error*

248 In survival analysis, Harrell's concordance index (Harrell Jr et al., 1982) is
 249 commonly used for estimating prediction error as it does not depend on choosing a
 250 fixed time for model evaluation and specifically accounts for censoring (May et al.,

251 2004). The concordance index (*C* index) is calculated by the following steps in ST
252 modelling.

- 253 1) Generate all possible pairs of samples over the data.
- 254 2) Remove pairs whose lower ST is censored. Remove pairs i and j if $st_i = st_j$ unless
255 at least one is an actual ST sample. The total number of permissible pairs is
256 recorded as *Per*.
- 257 3) For each permissible pair where $st_i \neq st_j$: if the thinner ST has worse predicted
258 outcome (higher cumulative hazard value), count 1; ii) otherwise, count 0.5. For
259 each permissible pair where $st_i = st_j$ and both are actual ST samples: i) if
260 predicted outcomes are equal, count 1; ii) otherwise, count 0.5. For each
261 permissible pair where $st_i = st_j$ and not both, are actual ST samples: i) if the
262 actual ST sample has a worse predicted outcome, count 1; ii) otherwise, count
263 0.5. The sum of all permissible pairs is recorded as *Con*.
- 264 4) The *C* index is defined by the ratio of *Con* to *Per*.

265 In RSF, the *C* index is computed via OOB data using the steps mentioned
266 above, and it ranges between 0 and 1. The prediction error is calculated by the 1-*C*
267 index, so it is also between 0 and 1. A lower prediction error represents better model
268 performance for the calibration model.

269

270 2.3.3. Assessing the main controlling factors for ST modelling

271 To assess the main controlling factors for ST in France, the variable
272 importance of the ST predictors (i.e., covariates used) in the RSF model were
273 evaluated. In RSF, the variable importance of a covariate c is calculated by dropping
274 OOB samples down their in-bag survival tree. A sub-node is randomly assigned
275 when encountering a split for c , and then an average of the CHF obtained from these

276 trees is calculated. The variable importance for c is calculated as the difference of
277 prediction error between the new ensemble obtained using randomized c
278 assignments and the original ensemble. A larger variable importance value indicates
279 a higher contribution to the model for a covariate.

280

281 2.4. Soil thickness probability mapping and bootstrapping for determining prediction 282 uncertainty

283 As introduced in Section 2.3.2, the RSF model outcome entails a function
284 between the survival (censored) probability and ST for each prediction. In other
285 words, the censored probability can be calculated over the full soil profile (0 to the
286 maximum depth of actual ST samples) for any position in mainland France from RSF.
287 As an example, the censored probabilities for the six *GlobalSoilMap* standard depths
288 were extracted from the survival probability function (Fig. 3); those depths are 5, 15,
289 30, 60, 100, and 200 cm, which we refer to hereafter as ST5, ST15, ST30, ST60,
290 ST100 and ST200, respectively. From this, we derived a probability map for each
291 *GlobalSoilMap* standard depth in mainland France.

292 Bootstrapping was applied to determine the average and 90% Confidence
293 Intervals (CIs) of the RSF model. Hence, we did not determine the 90% PIs as is
294 recommended by the *GlobalSoilMap* specifications; instead we estimated the 90%
295 CIs. This was deemed suitable, as we were not able to identify the random error in
296 the RSF model. Consequently, the estimated 90% CIs would be narrower than 90%
297 PIs. The bootstrap samples were drawn 50 times by repeated random sampling with
298 replacement of the RMQS sites; the RMQS sites not used in each bootstrap sample
299 were used to evaluate the model performance of each bootstrap RSF model (details
300 in Section 2.5). Note that the bootstrap sample used here corresponds to the initial

301 data used in the RSF framework (Fig. 2), not the bootstrap sample used to generate
302 trees. Finally, using these bootstrap samples, 50 bootstrap RSF models were
303 generated, from which 50 probability functions between the censored probability and
304 ST could be exhaustively predicted for mainland France. After several iterative model
305 calibrations leading to the final prediction model, we choose 50 bootstrap models
306 because it is time-consuming to make predictions at a 90 m resolution for mainland
307 France (RSF produces a probability function rather than a value for each pixel, so it
308 takes 2 weeks for 50 bootstrap RSF models under parallel computing that make full
309 use of a computer with 8 cores and 32 GB of RAM). A robust estimate of the
310 probability of exceeding each standard *GlobalSoilMap* soil depth was determined by
311 averaging the bootstrap predictions. Their lower and upper 90% CIs were calculated
312 by the averaged bootstrap predictions minus and plus 1.645 times (Z score for 90%
313 CIs) the standard deviation of bootstrap predictions, respectively. Surface area
314 percentages of five probability intervals (0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1)
315 were calculated from the averaged bootstrap predictions of probability maps at six
316 *GlobalSoilMap* standard depths. The mean probability was computed by averaging
317 all pixels of the probability map for each *GlobalSoilMap* standard depth.

318

319 2.5. Model performance

320 In addition to the CIs, the model performance of each *GlobalSoilMap* standard
321 depth was evaluated using the RMQS sites that were not used in the bootstrap
322 samples, which referred to an evaluation dataset from each bootstrap RSF model.
323 For a given *GlobalSoilMap* standard depth (st_s), the prediction performance was
324 evaluated based on the confusion matrix in which the misclassification rate was
325 calculated based on whether the data was censored or not. Hence, given a sample

326 with observed ST (st_o): 1) when $st_s \leq st_o$, if the probability exceeds 0.5, the sample is
327 correctly predicted, otherwise, it is incorrectly predicted; and 2) when $st_s > st_o$, if the
328 probability is less than 0.5, the sample is correctly predicted, otherwise, it is
329 incorrectly predicted.

330 Subsequently, the confusion matrix was calculated as the mean counts of
331 OOB samples with actual ST and censored ST separately from 50 bootstrap
332 predictions.

333 All of the statistics and modelling were performed in R (R Core Team, 2016). R
334 package *randomForestSRC* was used for RSF modelling (Ishwaran and Kogalur,
335 2017).

336

337 **3. Results**

338 3.1. Summary statistics of the ST dataset

339 Among 2108 RMQS sites, more than half were right censored for ST (Fig. 4).
340 The actual ST ranged from 0 to 300 cm, with a mean value of 64 cm. The first
341 quantile, median and third quantile were 39, 59, and 80 cm, respectively, indicating a
342 large percentage of soils thinner than 60 cm. The censored ST ranged from 50 to 270
343 cm, with the mean ST (104 cm) being higher than the actual observed RMQS sites.
344 For the censored RMQS sites, the first quantile, median, and third quantile were 75,
345 95, and 120 cm, respectively.

346

347 3.2. Model performance

348 The prediction error of the calibrated RSF models decreased from 0.27 to 0.15
349 as the number of trees increased up to 50 (Fig. 5). After 50 trees, the prediction error
350 decreased slightly and became more stable as the number of trees increased (max.

351 300 trees). This indicated that 50 trees were sufficient for this study to produce a
352 stable model while accelerating the prediction efficiency for big data.

353 The prediction performance differed when evaluated at the six *GlobalSoilMap*
354 standard depths (Table 2). For the actual RMQS sites, the overall accuracy
355 decreased from 0.989 to 0.546, when depth increased from ST5 to ST60. The overall
356 accuracy then gradually increased up to 0.793 for ST200. The overall accuracy for
357 censored RMQS sites were 1, 0.998, and 0.995, respectively, for SD5, SD15, and
358 SD30, the accuracy then decreased to 0.825 for SD60 and subsequently dropped to
359 0.534 for ST100 and 0.563 for ST200.

360

361 3.3. Controlling factors of ST modelling

362 Parent material (PM) and climatic zones (TYPO) were the two most important
363 variables affecting the ST probabilities in RSF models, based on the average
364 bootstrap RSF (Fig. 6). The difference in prediction error between the new and the
365 original ensembles was most affected by these two variables, despite the large 90%
366 CIs. Roughness, precipitation, elevation, slope, gravimetry and Net Primary
367 Production (NPP) also had large contributions in ST modelling. The remaining
368 covariates contributed less to the RSF model and had smaller CIs.

369

370 3.4. ST probability maps and associated confidence intervals

371 Fig. 7 presents the ST probability maps of exceeding the six *GlobalSoilMap*
372 standard depths and their 90% CIs for mainland France. Overall, the average
373 probability of exceeding the *GlobalSoilMap* standard depths of 5, 15, 30, 60, 100, and
374 200 cm were 0.99, 0.97, 0.88, 0.68, 0.51, and 0.42, respectively.

375 The probability of exceeding ST5 was close to 1 across the whole country,
376 except for eastern (the Alps) and southwestern France (the Pyrenees). The 90% CI
377 was very narrow (0.02 ± 0.06), indicating low model uncertainty and thus robust
378 estimates for the ST5 map.

379 A similar spatial distribution was observed when ST increased to ST15. The
380 low probability in southern France (the Massif Central) showed that this region had a
381 high probability of having STs less than 15 cm. The difference between the lower and
382 upper limits of the 90% CI was still low (0.05 ± 0.08), indicating a robust estimate.
383 Moreover, the surface area percentages for the five probability intervals were also
384 quite close to those of ST5 (Fig. 8).

385 When the ST depth criteria was further increased to ST30, in addition to
386 previously mentioned locations, low probabilities were found in eastern France (the
387 Jura Mountains, Fig. 7). Moreover, the CIs substantially increased (give numbers)
388 compared to ST5 and ST15. This indicates a larger prediction uncertainty and a
389 lower model robustness. In comparison with ST15, a slight increase (2%) was
390 observed for the surface area with probabilities between 0.4 and 0.6. The surface
391 area having a probability between 0.6 and 0.8 increased from 1 to 14%, while the
392 area with a probability between 0.8 and 1 decreased from 98 to 83% (Fig. 8).

393 Moving from the ST30 up to the ST200 thickness criteria, substantial changes
394 in spatial patterns and the probability of surpassing the ST criteria became apparent.
395 Most notable is how the surface area with probabilities between 0.8 and 1.0
396 continuously decreased, from 83% (ST30) to 2% (ST200). ST60 corresponded with a
397 probability of 27% and ST100 with a probability of 7% (Fig. 8).

398 For ST200, more than 50% of the territory of mainland France had a low
399 probability (<0.4) of exceeding ST by 2 m, while less than 17% of the areas had a

400 high probability (>0.6) of exceeding the ST by 2 m (Fig. 7). The areas with a high
401 probability were mainly located in southwestern France (the Landes of Gascony),
402 central France (Sologne), and northern France (thick loess deposits).

403

404 **4. Discussion**

405 Several ways to perform probability mapping have been proposed in the
406 literature since the 1990s. For instance, Bell et al. (1994) applied discriminant
407 analysis with a maximum-likelihood classification function to map the soil drainage
408 probability in south-central Pennsylvania, USA. von Steiger et al. (1996) mapped the
409 probability of exceeding the maximum tolerable heavy metal concentrations by
410 Disjunctive Kriging in northeast Switzerland. Richer-de-Forges et al. (2017) used
411 Logistic Regression Kriging in probability mapping of iron pan presence in sandy
412 podzols in southwest France. The largest differences between the methods used in
413 previous studies and RSF can be summarized in two aspects: 1) RSF is able to deal
414 with right censored data while others are not, and 2) RSF can potentially produce
415 probability estimates of any ST value, whereas other methods deal with
416 presence/absence at a given threshold for the soil attributes of interest. Moreover,
417 others used multiple sequential indicator simulations to model this type of distribution
418 (e.g., Cattle et al., 2002). The survival analysis we used is a similar approach, except
419 that it models the survival function rather than the empirical distribution function.

420 In the RMQS dataset, right censored ST observations entail more than half of
421 the observations. Using them for ST modelling with traditional DSM approaches
422 would result in highly underestimated ST estimates. Lacoste et al. (2016) proposed
423 adding a fixed value of 30 cm to censored samples before modelling, which may help
424 lower the underestimation but does not really solve the problem. Moreover, as actual

425 ST values of these censored sites remain unknown, adding a fixed value may even
426 add more noise to the data, and thus enlarging the prediction uncertainty. As shown
427 in Fig. 9, the probability of exceeding the observed ST for each censored RMQS site
428 was mainly between 0.5 and 1, with a median value of 0.78. Thereafter, as outlined
429 in the methodology Section 2.3.2, RSF makes use of the probability function derived
430 from right censored information, thereby avoiding underestimating ST at these
431 censored positions.

432 The mean probability of exceeding an ST of 100 cm across mainland France
433 was 0.51 (Fig. 8), which means that all locations have a 50% possibility of being
434 observed with an ST thicker than 100 cm. This result implies that the median ST in
435 mainland France is approximately 100 cm, which is in line with previous work by
436 Lacoste et al. (2016), showing that 48 and 54% of surface areas were below 100 cm
437 when using Gradient Boosting Modelling and Cubist models, respectively.

438 The results showed that the prediction performance decreased from 0 to 60
439 cm and subsequently increased up to 200 cm for censored RMQS sites (Table 2),
440 implying that the predicted probability of exceeding a given ST from the RSF model is
441 more reliable for extreme values (i.e., a thin ST or thick ST). Indeed, due to the soil
442 forming conditions in mainland France, except for steep slopes in mountainous areas,
443 very thin soils are quite rare, and thus the probability of exceeding a very thin ST is
444 high. Conversely, very thick soils are concentrated in (former) depositional areas
445 (valleys, aeolian sand, or loess deposits) that can be easily mapped using some of
446 the covariates (e.g., parent material and terrain parameters). For the censored
447 RMQS sites, the overall accuracies for ST100 and ST200 were approximately 0.5, in
448 which a large percentage of thin ST samples were misclassified as being thick. This
449 can be explained by the fact that we used observed ST of censored RMQS sites in

450 calculating the confusion matrix. Consequently, we may overestimate the percentage
451 of misclassification mentioned before and thus underestimate the overall accuracies
452 of ST100 and ST200.

453 Parent material and climatic zones were the most important variables for
454 predicting ST in France using the bootstrapped RSF, but roughness, precipitation,
455 elevation, slope, gravimetry, and NPP also substantially contributed to the ST model.
456 These results are in line with previous findings reported by Lacoste et al. (2016).
457 Lacoste et al. (2016) stated that the most important covariates of ST modelling in
458 mainland France were soil properties, climate covariates and land use. Considering
459 the variable importance and the variables acting as controlling factors for ST, parent
460 material, climatic zones, precipitation, and gravimetry are direct drivers for the
461 weathering process. Roughness, elevation, slope, and NPP are more related to
462 sediment transport dynamics.

463 Future research should aim to derive an ST map using RSF, instead of the
464 currently presented ST probability map of exceeding a given depth. There are three
465 ways to determine the actual soil ST from the unique probability function produced by
466 RSF for each location of interest: 1) use the ST extracted from the median probability
467 in the predicted function; 2) use the ST extracted from a fixed probability, allowing the
468 classification of censored and actual ST at high accuracy among RSF calibration
469 datasets; 3) perform a derivative analysis on the probability curve. Moreover, it will be
470 interesting to combine RSF with geostatistical methods. For example, kriging of
471 residuals (Hengl et al., 2004) that are not captured by RSF and/or sampling
472 optimizing for future campaigns to reduce the prediction variance at locations where it
473 is highest. Alternatively, the presented probability maps can be used directly for
474 additional ST sampling campaigns, aimed at ST modelling in mainland France. For

475 example, the regions with a high probability (>0.8) of ST200 have a large chance of
476 being censored. Integrating those high probability regions with parent material and
477 climatic zones would yield an efficient and effective sampling design using
478 conditioned Latin hypercube sampling (cLHS, Minasny and McBratney, 2006) to
479 obtain more representative samples of all physiographic contexts. RSF is able to
480 provide a probability at any depth and thus will be helpful for decision making in
481 geotechnical engineering regarding, for example, laying out drains, pipes, and tubes
482 (Zhang et al., 2005).

483 **5. Conclusions**

484 This study introduced the use of RSF in ST probability modelling to deal with
485 right censored data for Digital Soil Mapping. RSF produced a probability function of
486 ST for each soil sample included in the database. This function allowed the
487 estimation of a probability of exceeding a given ST, indicating each soil location was
488 right censored or not. Robust estimates were made by bootstrapping the RSF model
489 to quantify an averaged bootstrap prediction and 90% CI for each *GlobalSoilMap*
490 standard depth (5, 15, 30, 60, 100 and 200 cm) using the RSF survival probability
491 functions. The model evaluation indicated an overall good performance (overall
492 accuracy from 0.546 to 0.989) of RSF to predict the probability exceeding the six
493 *GlobalSoilMap* standard depths. The RSF proved suitable for using right censored
494 soil data for digital soil mapping, and thereby this work introduced a new approach
495 capable of using both right censored and actual data for modelling ST accordingly.

496

497 **Acknowledgement**

498 The collection and handling of soil data was supported by the French Scientific
499 Group of Interest on soils (GIS Sol), including the French Ministry of Ecology, the

500 French Ministry of Agriculture, the French Environment and Energy Management
501 Agency (ADEME), the French Institute for Research and Development (IRD), the
502 French Institute for National Geographic and Forest Information (IGN) and the
503 National Institute for Agronomic Research (INRA). We thank all the people involved
504 in sampling the sites and populating the database. We would also like to thank
505 Quentin Styc and Philippe Lagacherie for sharing their ideas on dealing with right
506 censored data. Songchao Chen received the support of the China Scholarship
507 Council for three years' of Ph.D. study in INRA and Agrocampus Ouest (under grant
508 agreement no. 201606320211).

509

510 **References**

- 511 Achache, J., Debeglia, N., Grandjean, G., Guillen, A., Le Bel, L., Ledru., P., Renaud,
512 X., Autran,A., Bonijoly, D., Calcagno, P., Pluchery, E., Guennoc, P., Truffert, C.,
513 Rossi, P., Vairon, J.,Avouac, J.P., Poli, E., Senechal, G., Brun, J.P., Galdeano,
514 A., Diament, M., Tarits, P., Mervier, J., Paul, A., Poupinet, G., Marquis, G., Bayer,
515 R., Chautra, J.M., 1997.GEOFRANCE 3D: l'imagerie geologique et geophysique
516 3D du sous-sol de la France. Mémoires de la Société Géologique de France,
517 172, 53–71.
- 518 Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a
519 large sample study. The Annals of Statistics 10(4), 1100–1120.
- 520 Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong,
521 S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-
522 Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A.,
523 Thompson, J.A., Zhang, G.-L., 2014a. Chapter Three — GlobalSoilMap: Toward
524 a Fine-Resolution Global Grid of Soil Properties. Adv. Agron. 125, 93–134.

525 Arrouays, D., McKenzie, N.J., Hempel, J., Richer de Forges, A.C., McBratney, A.,
526 2014b. GlobalSoilMap: Basis of the Global Spatial Soil Information System. 1st
527 ed. CRC Press Taylor & Francis Group.

528 Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.*
529 47(2), 151–163.

530 Bell, J.C., Cunningham, R.L., Havens, M.W., 1994. Soil drainage class probability
531 mapping using a soil-landscape model. *Soil Sci. Soc. Am. J.* 58(2), 464–470.

532 Bonfatti, B.R., Hartemink, A.E., Vanwalleghem, T., Minasny, B., Giasson, E., 2018. A
533 mechanistic model to predict soil thickness in a valley area of Rio Grande do Sul,
534 Brazil. *Geoderma* 309, 17–31.

535 Bouma, J., 2018. The challenge of soil science meeting society's demands in a “post-
536 truth”, “fact free” world. *Geoderma* 310, 22-28.

537 Bourennane, H., King, D., Chery, P., Bruand, A., 1996. Improving the kriging of a soil
538 variable using slope gradient as external drift. *Euro. J. Soil Sci.* 47(4), 473–483.

539 Breiman, L., 2001. Random forests. *Mach. Learn.* 45(1), 5–32.

540 Cattle, J.A., McBratney, A., Minasny, B., 2002. Kriging method evaluation for
541 assessing the spatial distribution of urban soil lead contamination. *J. Environ.*
542 *Qual.* 31(5), 1576–1588.

543 Cerdan, O., Govers, G., Le Bissonnais, Y., Van Oost, K., Poesen, J., Saby, N., Gobin,
544 A., Vacca, A., Quinton, J., Auerswald, K., Klik, A., Kwaad, F.J.P.M., Raclot, D.,
545 Ionita, I., Rejman, J., Rousseva, S., Muxart, T., Roxo, M.J., Dostal, T., 2010.
546 Rates and spatial variations of soil erosion in Europe: A study based on erosion
547 plot data. *Geomorphology* 122, 167–177.

548 Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Saby, N.P.
549 and Walter, C., 2019. National estimation of soil organic carbon storage potential

550 for arable soils: A data-driven approach coupled with carbon-landscape zones.
551 Sci. Total Environ. 666, 355–367.

552 Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D., 2018. Fine
553 resolution map of top-and subsoil carbon sequestration potential in France. Sci.
554 Total Environ. 630, 389–400.

555 Clothier, B.E., Hall, A.J., Deurer, M., Green, S.R., Mackay, A.D., 2011. Soil
556 ecosystem services: sustaining returns on investment into natural capital. In:
557 Sauer, T.J., Norman, J.M., Sivakumar, M.V.K. (Eds.), Sustaining Soil
558 Productivity in Response to Global Climate Change: Science, Policy, and Ethics.
559 Wiley-Blackwell, Oxford, pp. 117–139.

560 De Oliveira, V., 2005. Bayesian inference and prediction of Gaussian random fields
561 based on censored data. J. Comput. Graph. Stat. 14(1), 95–115.

562 Dietrich, W.E., Reiss, R., Hsu, M.L., Montgomery, D.R., 1995. A process - based
563 model for colluvial soil depth and shallow landsliding using digital elevation data.
564 Hydrol. Process. 9(3–4), 383–400.

565 Faroux, S., Tchuenté, A.K., Roujean, J.L., Masson, V., Martin, E., Le Moigne, P.,
566 2013. ECOCLIMAP-II/Europe: A twofold database of ecosystems and surface
567 parameters at 1 km resolution based on satellite information for use in land
568 surface, meteorological and climate models. Geosci. Model Dev. 6(2), 563–582.

569 Feranec, J., Jaffrain, G., Soukup, T., Hazeu, G., 2010. Determining changes and
570 flows in European landscapes 1990–2000 using Corine land cover data. Appl.
571 Geogr. 30(1), 19–35.

572 Fridley, B. L., Dixon, P., 2007. Data augmentation for a Bayesian spatial model
573 involving censored observations. Environmetrics 18(2), 107–123.

574 Groshans, G.R., Mikhailova, E.A., Post, C.J., Schlautman, M.A., 2018. Accounting for
575 soil inorganic carbon in the ecosystem services framework for United Nations
576 sustainable development goals. *Geoderma* 324, 37–46.

577 Harrell Jr, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the
578 yield of medical tests. *Jama* 247(18), 2543–2546.

579 Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1997. The soil production
580 function and landscape equilibrium. *Nature* 388, 358–361.

581 Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1999. Cosmogenic
582 nuclides, topography, and the spatial variation of soil depth. *Geomorphology*
583 27(1-2), 151–172.

584 Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 2001. Stochastic
585 processes of soil production and transport: Erosion rates, topographic variation
586 and cosmogenic nuclides in the Oregon Coast Range. *Earth Surf. Proc. Land.*
587 26(5), 531–552.

588 Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction
589 of soil variables based on regression-kriging. *Geoderma* 120(1–2), 75–93.

590 Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high
591 resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25,
592 1965–1978.

593 Info Terre – Site cartographique de référence sur les géosciences, 2014. Indice de
594 développement et de persistance des réseaux (IDPR), edited, BRGM – Centre
595 scientifique et technique, Orléans, France.

596 Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational
597 high resolution land cover map production at the country scale using satellite
598 image time series. *Remote Sens.* 9(1), 95.

599 Inventaire Forestier National, 2006. BD Forêt® In: Service de l'inventaire forestier et
600 statistique - Institut national de l'information géographique et forestière (IGN)
601 (Ed.), Nogent-sur-Vernisson, 613 France.

602 Ishwaran H., Kogalur U.B., 2017. Random Forests for Survival, Regression, and
603 Classification (RF-SRC). R package version 2.5.1.

604 Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival
605 forests. *The Annals of Applied Statistics* 2(3), 841–860.

606 IUSS Working Group, WRB, 2006. World reference base for soil resources. World
607 Soil Resources Report. 103.

608 Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled srtm for the globe
609 version 4. available from the CGIAR-CSI SRTM 90m Database.
610 <http://srtm.csi.cgiar.org>.

611 Jenny, H., 1941. *Factors of Soil Formation: A System of Quantitative Pedology*.
612 McGrawHill, New York, pp.1–20.

613 Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N., 2006. Le Réseau de
614 Mesures de la Qualité des Sols de France (RMQS). État d'avancement et
615 premiers résultats. *Étude et Gestion Sols* 13, 149–164.

616 Joly, D., Brossard, T., Cardot, H., Cavailhes, J., Hilal, M., Wavresky, P., 2010. Les
617 types de climats en France, une construction spatiale. *Cybergeo: European*
618 *Journal of Geography*.

619 Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete
620 observations. *J. Am. Stat. Assoc.* 53(282), 457–481.

621 Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A.,
622 Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett,

623 R.D., 2016. The significance of soils and soil science towards realization of the
624 United Nations Sustainable Development Goals. *Soil* 2(2), 111–128.

625 Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for
626 nationwide updating of the 1: 50,000 soil map of the Netherlands. *Geoderma*
627 241, 313–329.

628 King, D., Jones, R., Thomasson, A., 1995. *European Land Information Systems for*
629 *Agro-Environmental Monitoring*.

630 Knotters, M., Brus, D.J., Voshaar, J.O., 1995. A comparison of kriging, co-kriging and
631 kriging combined with regression for spatial interpolation of horizon depth with
632 censored observations. *Geoderma* 67(3–4), 227–246.

633 Kuriakose, S.L., Devkota, S., Rossiter, D.G., Jetten, V.G., 2009. Prediction of soil
634 depth using environmental variables in an anthropogenic landscape, a case
635 study in the Western Ghats of Kerala, India. *Catena* 79(1), 27–38.

636 Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016.
637 Evaluating large-extent spatial modeling approaches: A case study for soil depth
638 for France. *Geoderma Regional* 7(2), 137–152.

639 Leenaars, J.G., Claessens, L., Heuvelink, G.B., Hengl, T., González, M.R., van
640 Bussel, L.G., Guilpart, N., Yang, H., Cassman, K.G., 2018. Mapping rootable
641 depth and root zone plant-available water holding capacity of the soil of sub-
642 Saharan Africa. *Geoderma* 324, 18–36.

643 Marx, A., Erhard, M., Thober, S., Kumar, R., Schäfer, D., Samaniego, L., Zink, M.,
644 2019. Climate Change as Driver for Ecosystem Services Risk and Opportunities.
645 In *Atlas of Ecosystem Services*. Springer, Cham, pp. 173–178.

646 May, M., Royston, P., Egger, M., Justice, A.C., Sterne, J.A., 2004. Development and
647 validation of a prognostic model for survival time data: application to prognosis of

648 HIV positive patients treated with antiretroviral therapy. *Stat. Med.* 23(15), 2375–
649 2398.

650 McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping.
651 *Geoderma* 117(1–2), 3–52.

652 Meyer, M.D., North, M.P., Gray, A.N., Zald, H.S., 2007. Influence of soil thickness on
653 stand characteristics in a Sierra Nevada mixed-conifer forest. *Plant Soil* 294(1–2),
654 113–123.

655 Millennium Ecosystems Assessment. (2005). *Ecosystems and human well-being:
656 policy responses*. Island Press, Washington, DC.

657 Minasny, B., McBratney, A.B., 1999. A rudimentary mechanistic model for soil
658 production and landscape development. *Geoderma* 90(1–2), 3–21.

659 Minasny, B., McBratney, A. B., 2006. A conditioned Latin hypercube method for
660 sampling in the presence of ancillary information. *Comput. Geosci.* 32(9), 1378–
661 1388.

662 Mogensen, U.B., Ishwaran, H., Gerds, T.A., 2012. Evaluating random forests for
663 survival analysis using prediction error curves. *J. Stat. Softw.* 50(11), 1–23.

664 Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute
665 prediction using terrain analysis. *Soil Sci. Soc. Am. J.* 57(2), 443–452.

666 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016a.
667 GlobalSoilMap France: High-resolution spatial modelling the soils of France up
668 to two meter depth. *Sci. Total Environ.* 573, 1352–1369.

669 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016b.
670 National versus global modelling the 3D distribution of soil organic carbon in
671 mainland France. *Geoderma* 263, 16–34.

672 NASA LD, 2001. NASA Land Processes Distributed Active Archive Center (LP DAAC)
673 USGS/Earth Resources Observation and Science (EROS) Center.

674 Odeh, I.O.A., Chittleborough, D.J., McBratney, A.B., 1991. Elucidation of soil-
675 landform interrelationships by canonical ordination analysis. *Geoderma* 49(1–2),
676 1–32.

677 Odeh, I.O., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction
678 of soil properties from terrain attributes: heterotopic cokriging and regression-
679 kriging. *Geoderma* 67(3–4), 215–226.

680 Orton, T. G., Rawlins, B. G., Lark, R. M., 2009. Using measurements close to a
681 detection limit in a geostatistical case study to predict selenium concentration in
682 topsoil. *Geoderma* 152(3–4), 269–282.

683 Orton, T.G., Saby, N., Arrouays, D., Jolivet, C.C., Villanneau, E.J., Paroissien, J.B.,
684 Marchant, B.P., Caria, G., Barriuso, E., Bispo, A., Briand, O., 2012. Analyzing
685 the spatial distribution of PCB concentrations in soils Using below-quantification
686 limit data. *J. Environ. Qual.* 41(6), 1893–1905.

687 Pelletier, J.D., Rasmussen, C., 2009. Geomorphically based predictive mapping of
688 soil thickness in upland watersheds. *Water Resour. Res.* 45(9), 417.

689 Penížek, V., Borůvka, L., 2006. Soil depth prediction supported by primary terrain
690 attributes: a comparison of methods. *Plant Soil Environ.* 52(9), 424–430.

691 Rabot, E., Wiesmeier, M., Schlüter, S., Vogel, H.J., 2018. Soil structure as an
692 indicator of soil functions: a review. *Geoderma* 314, 122–137.

693 R Core Team. *R: A Language and Environment for Statistical Computing*. R
694 Foundation for Statistical Computing, Vienna, Austria (2016), pp. 1.

695 Richer-de-Forges, A.C., Saby, N.P., Mulder, V.L., Laroche, B., Arrouays, D., 2017.
696 Probability mapping of iron pan presence in sandy podzols in South-West
697 France, using digital soil mapping. *Geoderma Regional* 9, 39–46.

698 Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N.P.A., Cousin, I., Martin,
699 M.P., 2019. Uncertainty assessment of *GlobalSoilMap* soil available water
700 capacity products: a French case study. *Geoderma* 344, 14–30.

701 Román, J.R., Roncero - Ramos, B., Chamizo, S., Rodríguez - Caballero, E. and
702 Cantón, Y., 2018. Restoring soil functions by means of cyanobacteria inoculation:
703 importance of soil conditions and species selection. *Land Degrad. Dev.* 29(9),
704 3184–3193.

705 Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J.,
706 Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.dL.,
707 Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd,
708 K.D., Vågen, T.G., Vanlauwe, B., Walsh, M.G., Winowieki, L.A., Zhang, G.L.,
709 2009. Digital soil map of the world. *Science* 325 (5941), 680–681.

710 Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., Dai, Y., 2017. Mapping
711 the global depth to bedrock for land surface modeling. *J. Adv. in Model Earth Sy.*
712 9(1), 65–88.

713 Soil Survey Division Staff, 1993. Soil survey manual. Soil conservation service U.S.
714 Department of Agriculture Handbook, pp.18.

715 Styc, Q., Lagacherie, P., 2016. Predicting soil depth using a survival analysis model.
716 Oral presentation at the 7th Global Workshop on Digital Soil Mapping, Aarhus,
717 Denmark, 27 June to 1 July.

718 Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil
719 depth from topographic and land cover attributes. *Water Resour. Res.* 45(10),
720 438.

721 Vanwalleghem, T., Poesen, J., McBratney, A., Deckers, J., 2010. Spatial variability of
722 soil horizon depth in natural loess-derived soils. *Geoderma* 157(1–2), 37–45.

723 Vaysse, K., Lagacherie, P., 2015. Evaluating digital soil mapping approaches for
724 mapping GlobalSoilMap soil properties from legacy data in Languedoc-
725 Roussillon (France). *Geoderma Regional* 4, 20–30.

726 Vogel, H.J., Bartke, S., Daedlow, K., Helming, K., Kögel-Knabner, I., Lang, B., Rabot,
727 E., Russell, D., Stöbel, B., Weller, U., Wiesmeier, M., 2018. A systemic approach
728 for modeling soil functions. *Soil* 4(1), 83–92.

729 Von Steiger, B., Webster, R., Schulin, R., Lehmann, R., 1996. Mapping heavy metals
730 in polluted soil by disjunctive kriging. *Environmental Pollution* 94(2), 205–215.

731 Wang, J., Endreny, T.A., Hassett, J.M., 2006. Power function decay of hydraulic
732 conductivity for a TOPMODEL - based infiltration routine. *Hydrol. Process.*
733 20(18), 3825–3834.

734 Zhang, T., Frauenfeld, O.W., Serreze, M.C., Etringer, A., Oelke, C., McCreight, J.,
735 Barry, R.G., Gilichinsky, D., Yang, D., Ye, H., Ling, F., 2005. Spatial and
736 temporal variability in active layer thickness over the Russian Arctic drainage
737 basin. *Journal of Geophysical Research: Atmospheres* 110, D16101.

738 Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using
739 GIS, expert knowledge, and fuzzy logic. *Soil Sci. Soc. Am. J.* 65(5), 1463–1472.

740

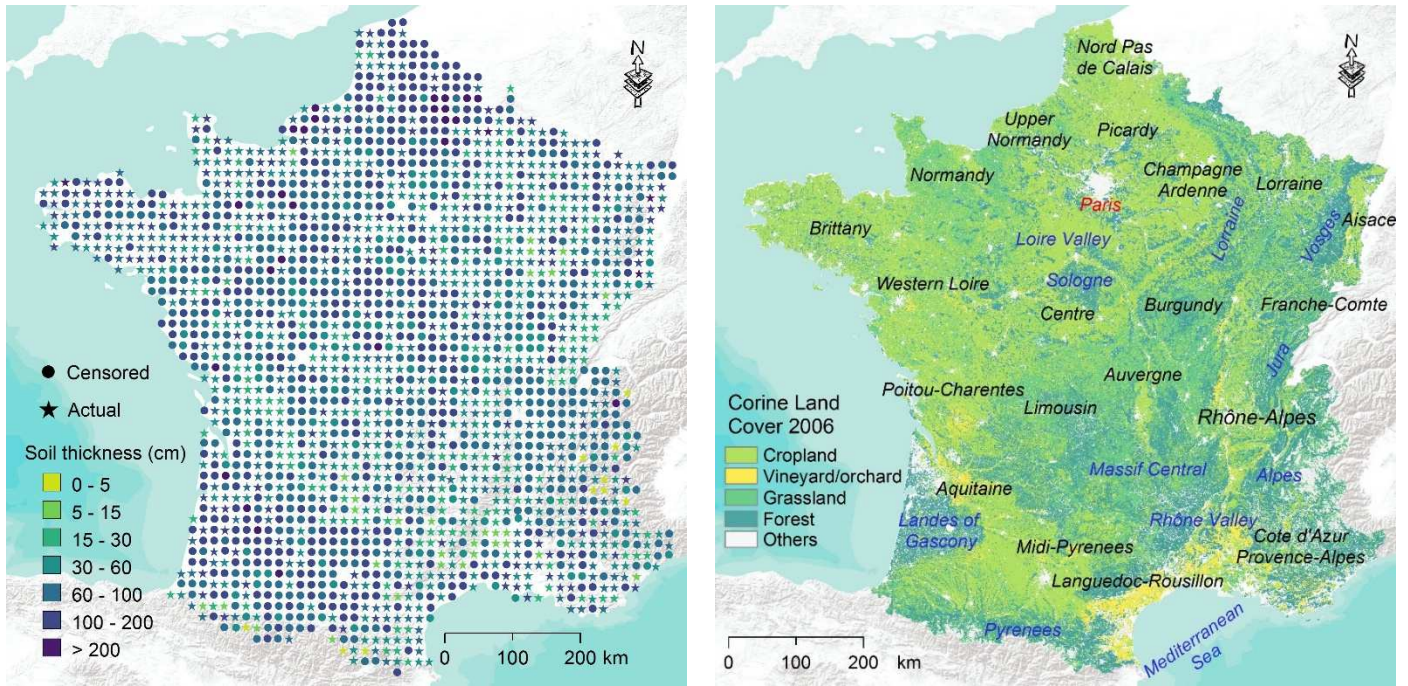
741 **Figures**

742 Fig. 1 Locations of RMQS sites with actual (dotted) and censored (star) ST values.

743 For each site, ST is classified based on the *GlobalSoilMap* standard depths. Corine

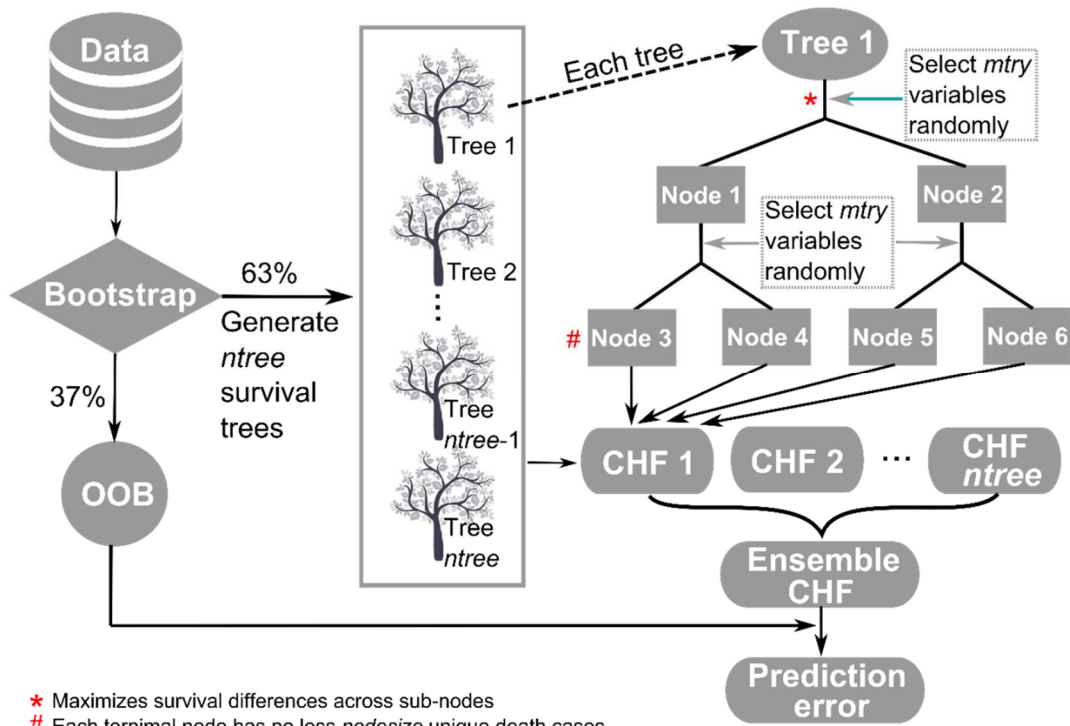
744 Land Cover map of 2006 of mainland France (right) with the administrative regions

745 (black italics) and natural geographic regions (blue italics).



746

747 Fig. 2 Random survival forest workflow.



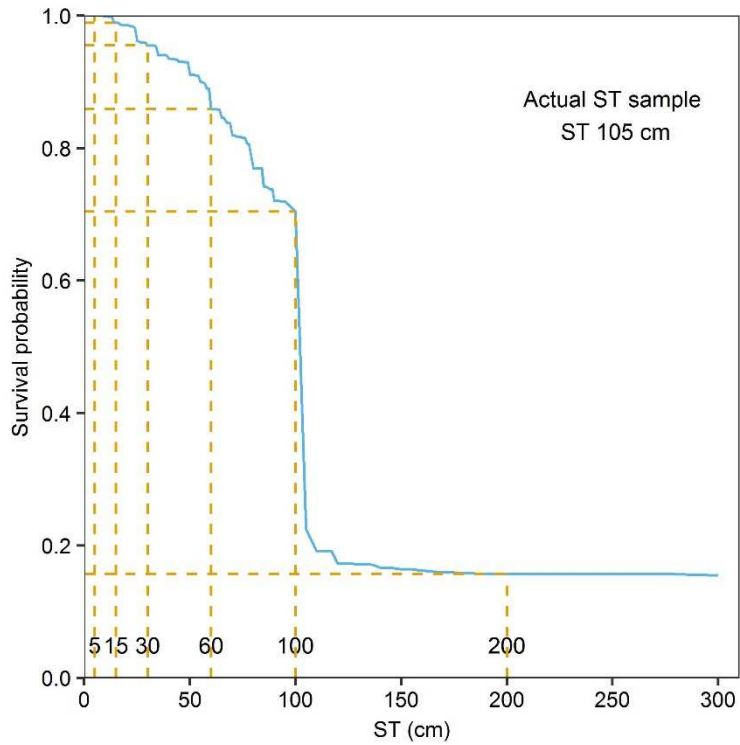
* Maximizes survival differences across sub-nodes

Each terminal node has no less $nodesize$ unique death cases

748

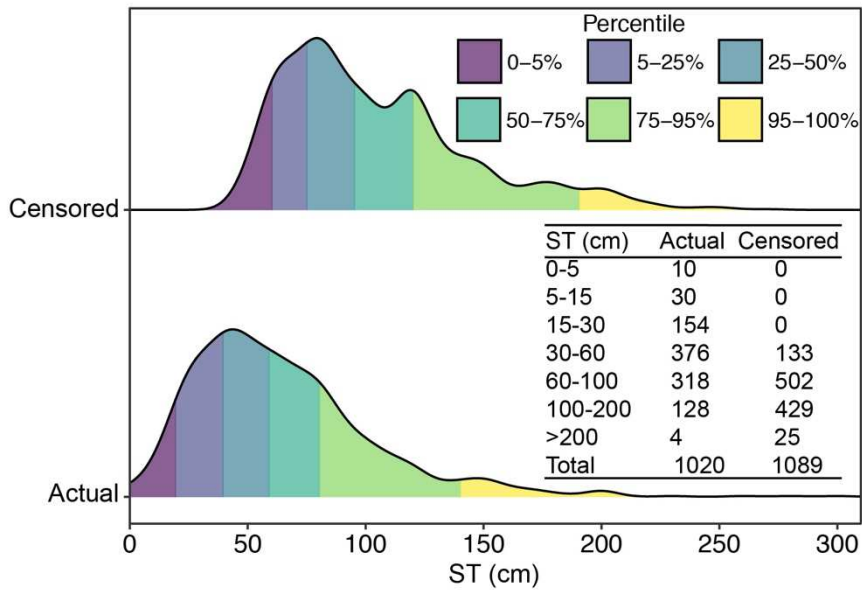
749

750 Fig. 3 Survival probability curve (blue solid line) for one location predicted by RSF.
751 The orange dashed vertical lines indicate the six *GlobalSoilMap* standard depths,
752 and the orange dashed horizontal lines indicate their corresponding censored
753 probabilities that are derived from the survival probability curve.



754
755

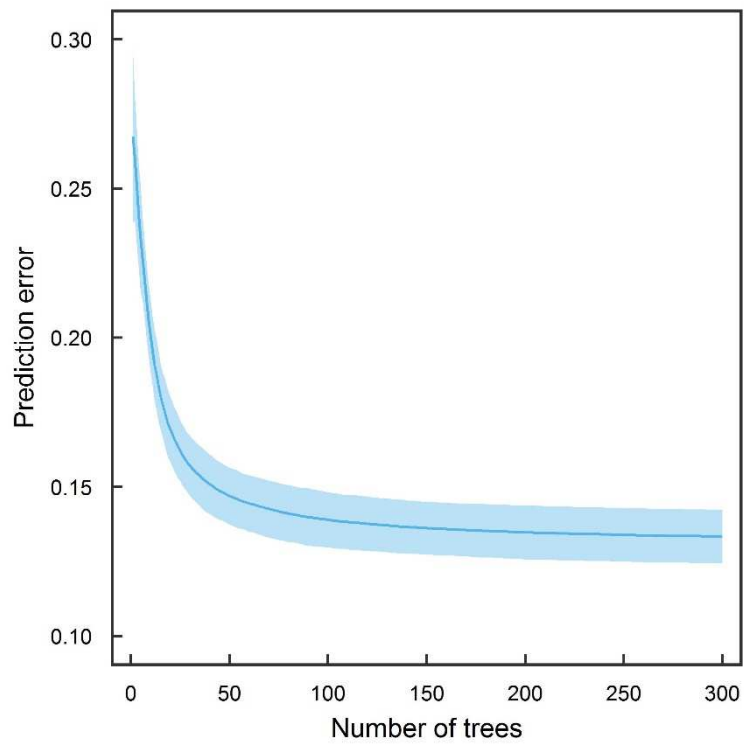
756 Fig. 4 Density distribution of STs for actual and censored RMQS sites. Counts of
 757 actual and censored samples within *GlobalSoilMap* depth intervals are provided.



758

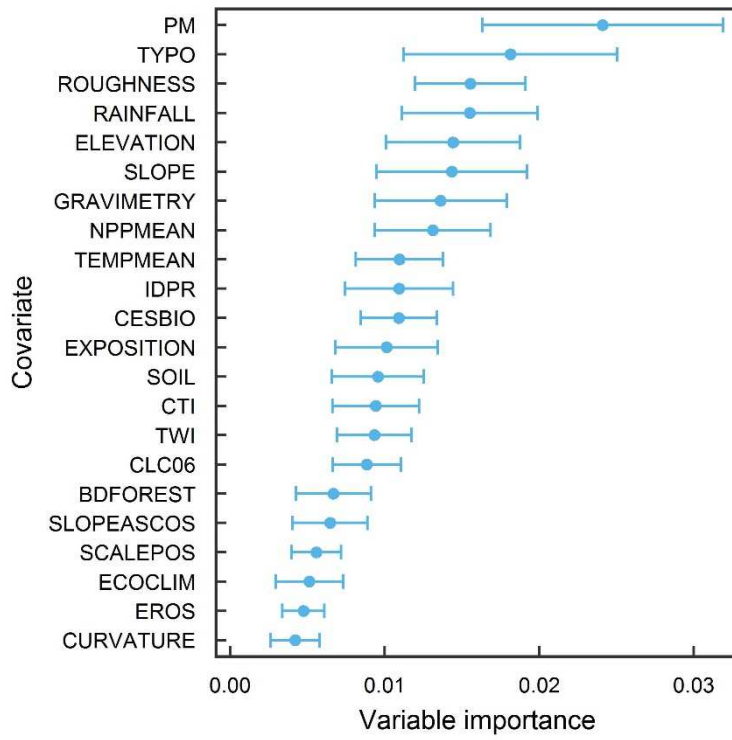
759

760 Fig. 5 Mean and 90% confidence intervals of the prediction error, given different
761 numbers of trees from 50 bootstrapping random survival forests.



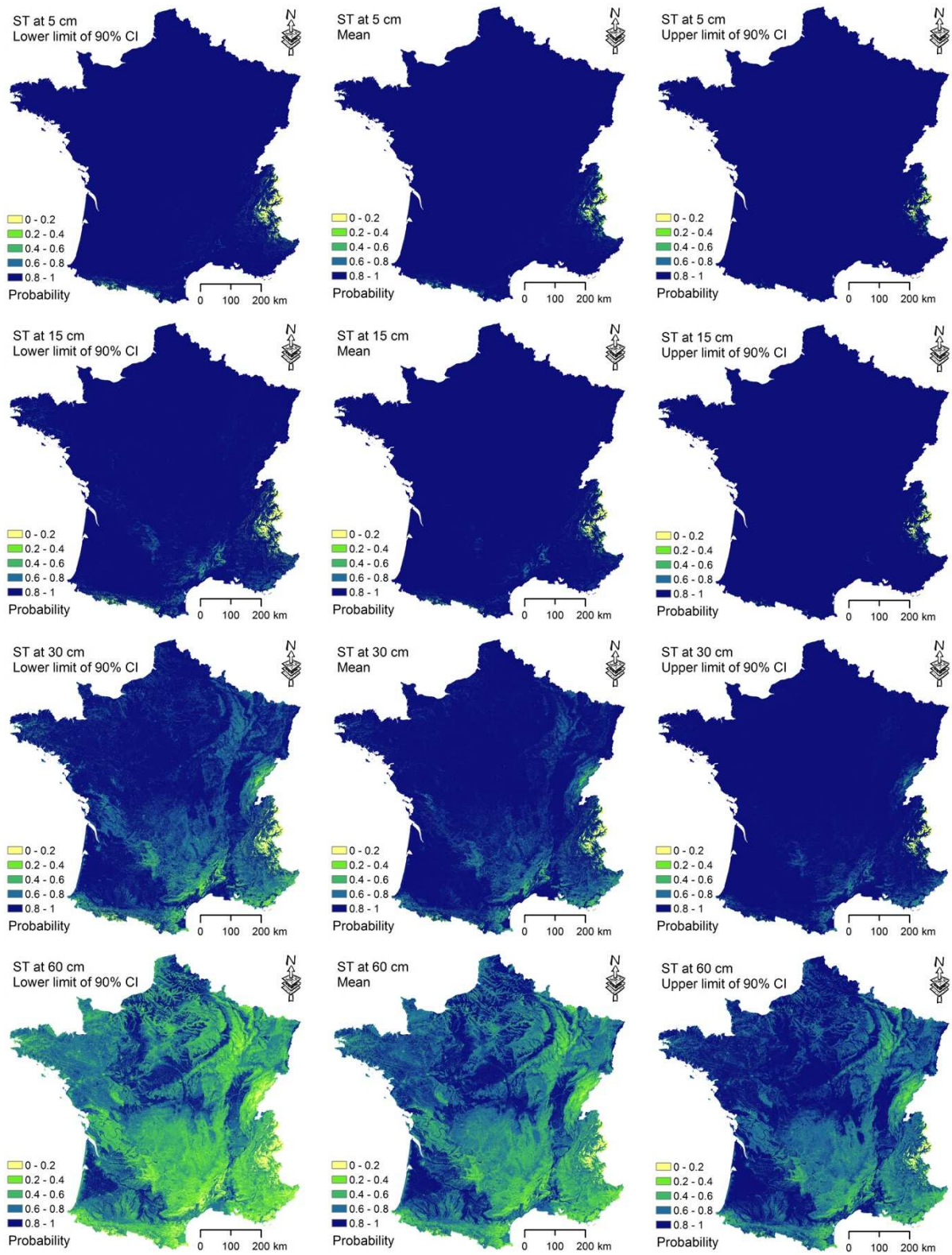
762

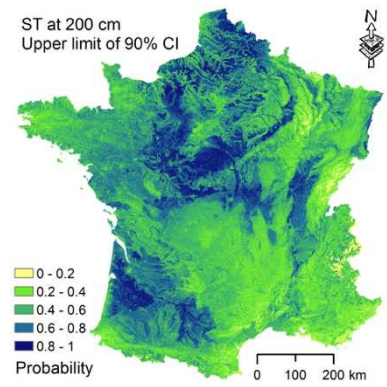
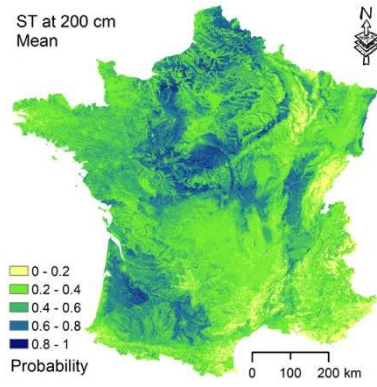
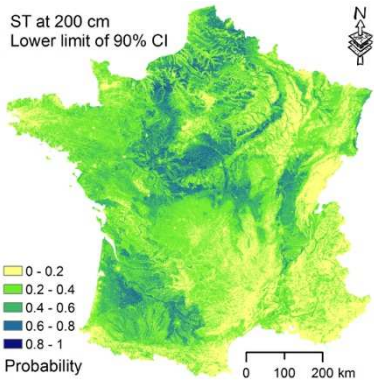
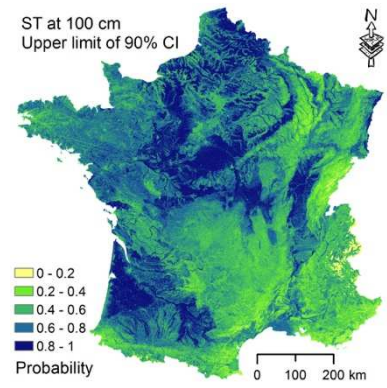
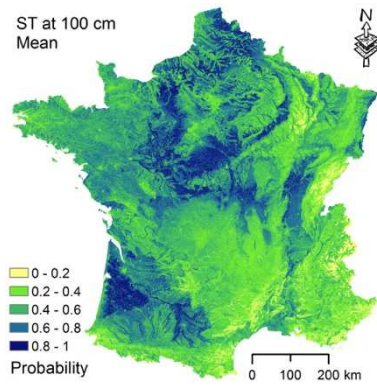
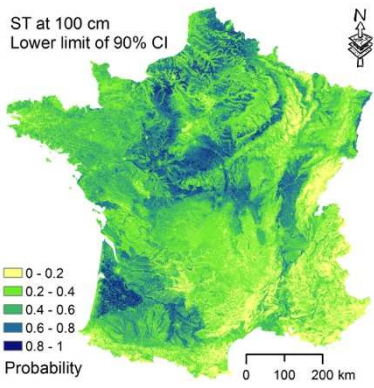
763 Fig. 6 Mean and 90% confidence intervals of variable importance from 50
764 bootstrapping random survival forests.



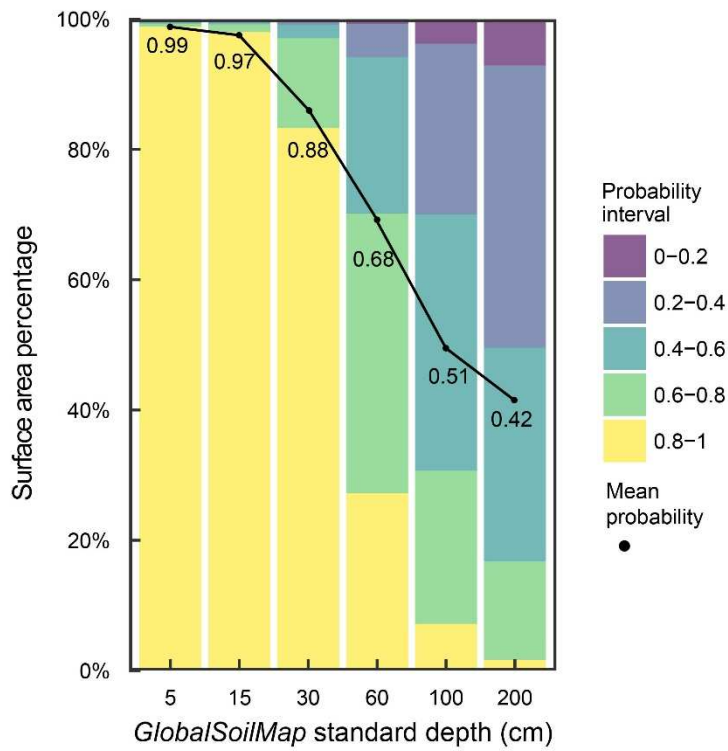
765

766 Fig. 7 Probability maps of exceeding the six *GlobalSoilMap* standard depths (middle)
767 and their associated 90% confidence intervals (left and right).



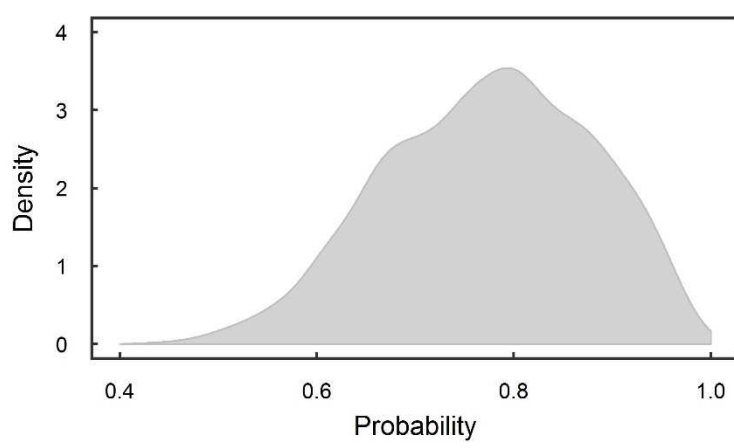


768 Fig. 8 Surface area percentage of the probability of exceeding the ST of each
 769 *GlobalSoilMap* standard depths. The mean probability is calculated by averaging all
 770 the pixels in the probability map for each *GlobalSoilMap* standard depth.



771

772 Fig. 9 Histogram with the probability of exceeding the observed ST for each censored
773 RMQS site.



774
775

777 **Table 1 Exhaustive covariates used for ST modelling (after Mulder et al., 2016b)**

Variable	Abbreviation	Scale/resolution	Soil forming factor	Reference
Elevation	ELEVATION	90 m	Relief	Jarvis et al. (2008)
Compound topographic index	CTI	90 m	Relief	Jarvis et al. (2008)
Curvature	CURVATURE	90 m	Relief	Jarvis et al. (2008)
Exposition	EXPOSITION	90 m	Relief	Jarvis et al. (2008)
Roughness	ROUGHNESS	90 m	Relief	Jarvis et al. (2008)
Slope	SLOPE	90 m	Relief	Jarvis et al. (2008)
Slope cosines	SLOPECOS	90 m	Relief	Jarvis et al. (2008)
Slope position	SLOPEPOS	90 m	Relief	Jarvis et al. (2008)
Topographic wetness index	TWI	90 m	Relief	Jarvis et al. (2008)
Gravimetric data (Bouguer anomaly)	GREVIMETRY	4 km	Relief	Achache et al. (1997)
Soil type ^a	SOIL	1:1000000	Soil	IUSS Working Group WRB (2006)
Erosion rates	EROS	1:1000000	Soil	Cerdan et al. (2010)
Rate of river network development and persistence	IDPR	1:50000	Soil and parent material	Info Terre – Site cartographique de référence sur les géosciences (2014)
Parent material	PM	1:1000000	Parent material	King et al. (1995)
Mean annual net primary production	NPPMEAN	1 km	Organisms	NASA LD (2001)
Forest type	BDFOREST	Min area 2.25 ha	Organisms	Inventaire Forestier National (2006)
Land cover from Sentinel-2	LCS	10 m	Organisms	Inglada et al. (2017)
Corine land cover 2006	CLC06	250 m	Organisms	Feranec et al. (2010)
ECOCLIMAP land use	ECOCLIM	1 km	Organisms	Faroux et al. (2013)
Climatic zones	TYPO	1 km	Climate	Joly et al. (2010)
Mean annual precipitation	RAINFALL	1 km	Climate	Hijmans et al. (2005)
Mean annual temperature	TEMPMEAN	1 km	Climate	Hijmans et al. (2005)

778 ^a Soil type defined by World Reference Base (WRB)

779 Table 2 Model performance of actual and censored RMQS sites per each
 780 *GlobalSoilMap* standard depth, based on out of bag samples. The count of correctly
 781 classified sites is marked **bold**, and the overall accuracy is marked *italic underlined*.

ST (cm)	Predicted Observed	Actual RMQS sites			Censored RMQS sites		
		Thin	Thick	Accuracy	Thin	Thick	Accuracy
5	Thin	2	2	0.500	0	0	1
	Thick	0	367	1	0	400	1
	Reliability	1	0.995	<u>0.989</u>	<i>n.a.</i>	1	<u>1</u>
15	Thin	2	12	0.143	0	0	<i>n.a.</i>
	Thick	0	356	1	1	399	1
	Reliability	1	0.967	<u>0.962</u>	0	1	<u>0.998</u>
30	Thin	5	65	0.063	0	0	<i>n.a.</i>
	Thick	1	300	0.997	2	398	1
	Reliability	0.833	0.843	<u>0.822</u>	0	1	<u>0.995</u>
60	Thin	58	150	0.279	7	43	0.140
	Thick	18	144	0.889	27	323	0.923
	Reliability	0.763	0.490	<u>0.546</u>	0.205	0.883	<u>0.825</u>
100	Thin	203	120	0.628	97	138	0.413
	Thick	19	28	0.596	49	117	0.705
	Reliability	0.914	0.189	<u>0.624</u>	0.664	0.459	<u>0.534</u>
200	Thin	294	76	0.795	219	172	0.560
	Thick	1	1	0.500	3	6	0.667
	Reliability	0.997	0.013	<u>0.793</u>	0.986	0.034	<u>0.563</u>

782 *n.a.* Not available.