Uncertainty assessment of soil available water capacity using error propagation: A test in Languedoc-Roussillon

Quentin Styc, Philippe Lagacherie

Uncertainty assessment of soil available water capacity using error propagation: a test in Languedoc-Roussillon

Styc Quentin[1,2]

Lagacherie Philippe[1]

[1] LISAH, Univ Montpellier, INRAE, IRD, Montpellier SupAgro, Montpellier, France

[2] BRL Exploitation, Nîmes, France

Corresponding author : Philippe Lagacherie

Abstract

Soil available water capacity (SAWC) is a key soil indicator that plays a major role in many ecosystem services, such as food production, irrigation management, soil drought, flood control, and climate and gas regulation. Digital soil mapping (DSM) can be used to obtain needed SAWC maps. However, SAWC differs from the usual soil properties considered in DSM in that it involves several soil properties determined at several soil layers. Therefore, a specific approach is required to obtain SAWC maps and the associated uncertainty predictions.

The objective of this study was to build a SAWC mapping approach that could predict SAWC values at three maximum rooting depths (60, 100 and 200 cm) and their associated prediction uncertainties.

The approach was tested in the Languedoc-Roussillon region (southern France). Elementary available water capacities of each layers (in $cm.cm^{-1}$) and soil layer thicknesses were first mapped separately at 0-30, 30-60, 60-100 and 100-200 cm and then aggregated to estimate the SAWCs at the three mentioned maximum rooting depths. SAWC uncertainty was estimated with an error propagation model that used a first-order Taylor analysis. This analysis considered the mapping errors of each involved property, which were estimated by the quantile regression forest algorithm. We tested different error propagation models that differently considered the correlations between these mapping errors: no correlation considered, correlations between soil layer thicknesses and elementary water capacities per soil layer only, correlations between soil layers only, or all correlations considered.

The performances of both SAWC predictions and their uncertainties were assessed with a 10-fold cross validation that was iterated 20 times. The SAWC predictions showed poor accuracies (percentages of explained variance ranged from 0.12 to 0.13). The uncertainties of SAWC predictions were best estimated when the correlations between the soil layer errors were considered in the error propagation model whereas the uncertainties of SAWC predictions were severely underestimated when these correlations were neglected.

In spite of the poor performance in predicting SAWC at the punctual level due to the low density of soil observations (1/19 km^2), the SAWC approach appeared promising since it produced maps that agreed with the available pedological knowledge and precisely estimated the uncertainties.

Keywords: soil available water capacity, digital soil mapping, uncertainty, error propagation

Abbreviations: SAWC: soil available water capacity; DSM: digital soil mapping, AWC: available water capacity; AWC$_E$: elementary available water capacity; ST: soil thickness; MSOD: maximal soil observation depth; PTF: pedotransfer function; FC: soil water content at field capacity; PWP: soil water content at permanent wilting point.

1. Introduction

Soil available water capacity (SAWC) refers to the capacity of soils to store water for plants (Veihmayer and Hendrickson, 1927). SAWC is a key soil indicator that plays a major role in many ecosystem services, such as food production, irrigation management, soil drought, flood control, and climate and gas regulation. It is therefore a fundamental parameter that has been used in land evaluations and recently in soil ecosystem service assessments (Dominati et al., 2014). Currently, SAWC is operationally computed in the literature as follows (Cousin et al., 2003):

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \quad (1)$$

where SAWC is the soil available water capacity (cm), dh_i = the thickness of the *i*th horizon (cm), bd_i = the bulk density of the *i*th horizon ($g.cm^{-3}$), st_i = the coarse fragment content of the *i*th horizon (% volumetric), and θr_i and θw_i are the soil water contents at field capacity (FC) (i.e., the soil water content that remains in the soil after water has drained due to gravitational force) and at permanent wilting point (PWP) (i.e., the soil water retained so strongly that it is no longer available for plant roots, so plants wither and cannot recover their turgidity) of the *i*th horizon ($cm^3.cm^{-3}$), respectively.

To meet the need to map the SAWC, digital soil mapping (McBratney, 2003) can be considered an adequate approach since it provides the best solution for synergizing all data on the soils and its drivers that can be available in a given region, regardless of its size. Styc and Lagacherie (2019) proposed a modified equation for calculating SAWC (Eq. 2) to make it more easily mappable. The modifications consisted in i) replacing the difficult-to-measure FC and PWP inputs by pedotransfer functions that only involved available primary soil properties that are more current in the soil databases and ii) harmonizing the mapping inputs across sites by replacing the horizons defined by variable depth intervals by soil layers defined by fixed depth intervals.

$$SAWC_{map} = \sum_{i=1}^{n} SLT_i * \left(\frac{100-st_i}{100}\right) * \left[\left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon_{FC}\right) - \left(\sum_{j=1}^{n} \beta_j PP_j + \varepsilon_{PWP}\right)\right] \quad (2)$$

where SLT_i is the thickness of the soil layers fixed by the soil depth interval, st_i is the coarse fragment content of the ith horizon (% volumetric), $\alpha_1 \ldots \alpha_n$ and $\beta_1 \ldots \beta_n$ are the coefficients of the pedotransfer functions used to calculate the volumetric water contents at the field capacity and permanent wilting point respectively, $PP_1 \ldots PP_n$ are the values of the primary soil properties used as inputs for the pedotransfer functions (most often textural fractions) and ε_{FC}

and ε_{PWP} are the errors of the pedotransfer functions used for estimating Field capacity and Permanent wilting point respectively. Equation 2 shows that the SAWC determination involved several primary soil properties determined at several soil layers, which created methodological questions that have not been addressed by the classical DSM framework. First, there is no consensus about the inference trajectory selected for predicting SAWC. In the DSM literature, there were i) studies that calculate first AWC at the observed sites prior using these sites for calibrating a mapping function (Vanderlinden et al., 2005; Poggio et al., 2010; Hong et al., 2013) and ii) studies that mapped the AWC components first and then combined the mapping outputs to obtain an estimate of AWC (Ugbaje and Reuter, 2013, Leenaars et al., 2018, Román Dobarco et al., 2019). However, a comparison of 18 inference trajectories that combined different AWC calculations from the primary soil properties, soil layer aggregation and mapping (Styc and Lagacherie, 2019) showed significant differences in SAWC prediction accuracies, and none of the two inference trajectories cited above were optimal. The best inference trajectory was an intermediate trajectory that, before mapping, calculated the AWC for four soil layers.

Another methodological question was the ex-ante uncertainty assessment of the SAWC mapping output. In the classical DSM framework, the different models can provide a local estimate of the uncertainty of the predicted values of the target soil properties (Heuvelink et al, 2014, Vaysse & Lagacherie, 2017). Obtaining a similar estimate for the SAWC map requires an error propagation model that combines the different errors associated with the mapping of each layer of the soil properties involved in the SAWC calculation. Román Dobarco et al. (2019) used a first-order Taylor analysis to propagate mapping errors and pedotransfer function errors to the final SAWC predictions. They showed that the mapping of the SAWC components (soil texture and coarse fragment) was the main source of SAWC mapping uncertainty. However, the soil thickness mapping errors were not considered in their analysis, although Algayer et al. (2019) demonstrated that soil thickness can be the most critical component in the SAWC estimation. Furthermore, the error propagation model proposed by Román Dobarco et al. (2019) neglected the error correlations between the SAWC components, which assumed that these errors were independent of each other, which has not been demonstrated.

The objective of this study was to build a SAWC mapping approach with the best possible inference strategy that could predict all SAWC values for three maximum rooting depths (60, 100 and 200 cm) and their associated prediction uncertainties while taking into account all SAWC component mapping errors and their correlations. The approach was tested in the former Languedoc-Roussillon region (Southern France).

2. The Study Case

2.1. Study area

This study was carried out in the former Languedoc-Roussillon French administrative region, which is now part of the new Occitanie region (Figure 1). Located in southern France, the former region covers 27,236 km² of land that stretches from the Mediterranean Sea to the Pyrenees and Massif Central mountains. The region includes a wide-ranging diversity of climates, geologies, and landscapes that lead to a large pedodiversity, with 18 WRB major soil groups, which represent 75% of all soil groups in Europe, being included in this study area. Further details can be found in Vaysse and Lagacherie (2015, 2017).

Figure 1. Location of the study area (in color)

2.2. Soil profiles with observed SAWC components

In this study, we used a legacy dataset of 2024 measured soil profiles from Vaysse and Lagacherie (2015, 2017). SAWC for the different soil profiles were harmonized for providing AWC component values at the fixed soil layers that were considered for SAWC mapping (0-30 cm, 30-60 cm, 60-100 cm and 100-200 cm) (see explanation further on section 3.3.). For that, each AWC components were estimated at these layers from the initial soil horizons using mass conservation cubic splines (Bishop et al, 1999). For each considered soil layer, we selected the profiles at which all AWC components (soil texture, coarse fragments and soil layer thicknesses) were fully documented (see details below). This resulted in a reduction of the number of soil profiles for mapping each layer (Table 1).

Table 1. Number of soil profiles for each soil layer

Soil layer depths (cm)	Number of profiles
0-30	1464
30-60	1323
60-100	1064
100-200	822

Documenting soil layer thicknesses

The soil layer thicknesses (SLT) were documented by considering the following rules :

- If the lower limit of the soil layer (LL) was less than both the maximum soil observation depth (MSOD) of the soil profiles and the upper depth of a lithic or paralithic contact (UDPLC), then SLT was equal to the difference between its fixed lower and upper limits (e.g. the SLT of 30-60 cm soil layer is 30 cm)
- Else if LL was less than MSOD but greater than UDPLC, SLT was equal to the difference between UDPLC and LL (e.g. the SLT of the 30-60 cm layer with a lithic contact appearing at 50 cm is 20 cm)
- Else if LL was greater than MSOP, the SLT could not be determined, which lead to remove the soil layer from the input soil dataset.

2.3. Pedotransfer functions

In this study, we used the national-level pedotransfer functions (PTFs) developed by Román Dobarco et al. (2019) because our case study was in the domain of applicability of these PTFs, which ensured the best possible performances (Roman Dobarco et al, 2019). The volumetric soil water contents at field capacity (Eq. 3) and permanent wilting point (Eq. 4). These PTFs used clay and sand contents (%) as the predictive variables, which were calculated as follows:

$$\hat{\theta}_{ri} = 0.278 + 2.45\ 10^{-3}\ Clay - 1.35\ 10^{-3}\ Sand \tag{3}$$

$$\hat{\theta}_{wi} = 0.080 + 4.01\ 10^{-3}\ Clay - 2.93\ 10^{-4}\ Sand \tag{4}$$

where $\hat{\theta}_{ri}$ and $\hat{\theta}_{wi}$ are the volumetric water contents at field capacity and permanent wilting point, respectively.

2.4. Soil covariates

The employed DSM process, which relies on the scorpan model (McBratney et al., 2003), used the quantitative relationships between the target soil properties and available spatial variables related with soil, which are also called the "soil covariates".

The soil covariates of this study area were selected by Vaysse and Lagacherie (2015) following

two criteria: i) they could be derived from freely available geo-datasets for at least the French national level, and ii) they had a logical and process-based relationship with soil properties according to the literature. The soil covariates (table 1) accounted for the impact of topography, climate, organisms, and parent material. The regional-scale map (1:250,000) that delinated the major landscape types across the region was also considered as a soil covariate. All the soil covariates were computed at the nodes of the 90 m x 90 m grid of the SRTM Digital Elevation model, which corresponded also to the resolution of the predicted SAWC map. More details can be found in the descriptions of several applications of DSM to the region (Vaysse and Lagacherie, 2015, 2017; Styc and Lagacherie, 2019).

Table 2. The soil covariates

Variables	Abbreviation	Resolution/Scale	Source	Soil forming factor[1]	Type[2]
Topography					
Elevation	ELEV	90 m	SRTM	r	Q
Multiresolution Valley Bottom Flatness	MRVBF	90 m	SRTM	r	Q
Slope	SLOPE	90 m	SRTM	r	Q
Topographic Wetness Index	TWI	90 m	SRTM	r	Q
Plan Curvature	PLANCURV	90 m	SRTM	r	Q
Profile Curvature	PROCURV	90 m	SRTM	r	Q
Multiresolution Ridge Top Flatness	MRRTF	90 m	SRTM	r	Q
Topographic Position Index	TPI	90 m	SRTM	r	Q
Geology					
Hardness	HARDNESS	90 m	Geological map/soil profile	p	C
Texture	TEXTURE	90 m	Geological map/soil profile	p	C
Mineralogy	MINERALOGY	90 m	Geological map/soil profile	p	C
Climate					
Martonne Index	MARTONNE	90 m	WorldClim	c	Q
Emberger Index	EMBERGER	90 m	WorldClim	c	Q
Maximum Temperature	TMAX	90 m	WorldClim	c	Q
Minimum Temperature	TMIN	90 m	WorldClim	c	Q
Precipitation	PRECIPITATION	90 m	WorldClim	c	Q
Organisms					
Land Use	LANDUSE	30 m	Landsat 7	o	C
Soil					
Soil Map	SOILMAP	1:250 000	RRP	s	C

[1]: SCORPAN factors (s=soil property, c = climate, o = organisms, r = relief, p=parent material)

[2]: Q = quantitative, C = categorical

SRTM = Shuttle Radar Topography Mission; RRP = Référentiel Régional Pédologique.

3. The method

3.1. Random Forest

Random forest models (Breiman, 2001) are an ensemble learning method for both classification and regression. A forest, which is an ensemble of randomized decision trees, is built and trained based on a bootstrap approach. Individual trees are built using the principle of recursive partitioning. "*The feature space is recursively split into regions containing observations with similar response value*" (Strobl et al., 2009). The predictions of the individual trees are finally averaged to give one single prediction.

3.2. Mapping model: the quantile regression forest

In this study, we use one of the most commonly used algorithms in DSM studies, namely, the quantile regression forest algorithm (QRF; Meinshausen, 2006), which is an extension of Breiman's random forests (RF; 2001). For the regression, RF provides an ensemble prediction based on *n* regression trees. For every tree, the algorithm integrates random features by randomly selecting a subset of features to be split. While RF provides solely the conditional mean, QRF supplies the whole conditional distribution of the target variable by keeping all observations at the terminal nodes and can infer estimates for the conditional quantiles (Meinshausen, 2006). More details on QRF can be found in Meinshausen (2006).

QRF was run with the ranger package, which is a fast implementation of Breiman's random forest and Meinshausen's quantile regression forest (Wright and Ziegler, 2015).

3.3. Inference trajectories

Since SAWC is a soil indicator that involves several soil properties determined at several soil layers, it can be estimated following various possible inference following the order with which the three different steps, i.e., "combining primary soil properties", "aggregating soil layers across depths" and "mapping" are executed to provide the targeted output (Styc and Lagacherie, 2019). Styc and Lagacherie (2019) tested a total of 18 inference trajectories for throughout Languedoc-Roussillon that were performed to obtain the most appropriate SAWC map. From this study, we considered the best performing inference trajectory, i.e., we computed first AWC in four layers (0-30, 30-60, 60-100 and 100-200 cm) obtained by merging the three first layers defined in the GlobalSoilmap specifications (Arrouays et al, 2014), mapped them and then aggregating the maps of the four soil layers to obtain the final SAWC maps. To account for the different possible rooting depths across the different crops, these aggregations were performed over three different maximal rooting depth (60 cm, 100 cm, 200 cm).

However, we modified the inference trajectory (Figure 2) by mapping, the soil thickness and the elementary available water capacity (AWC_E) separately for each layer. AWC_E represent the water retention capacity for one centimeter of soil (in cm.cm^{-1}) and is defined as follows:

$$AWC_E = \left(\frac{100-st_i}{100}\right) * \left[\left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon\right) - \left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon\right)\right] \quad (5)$$

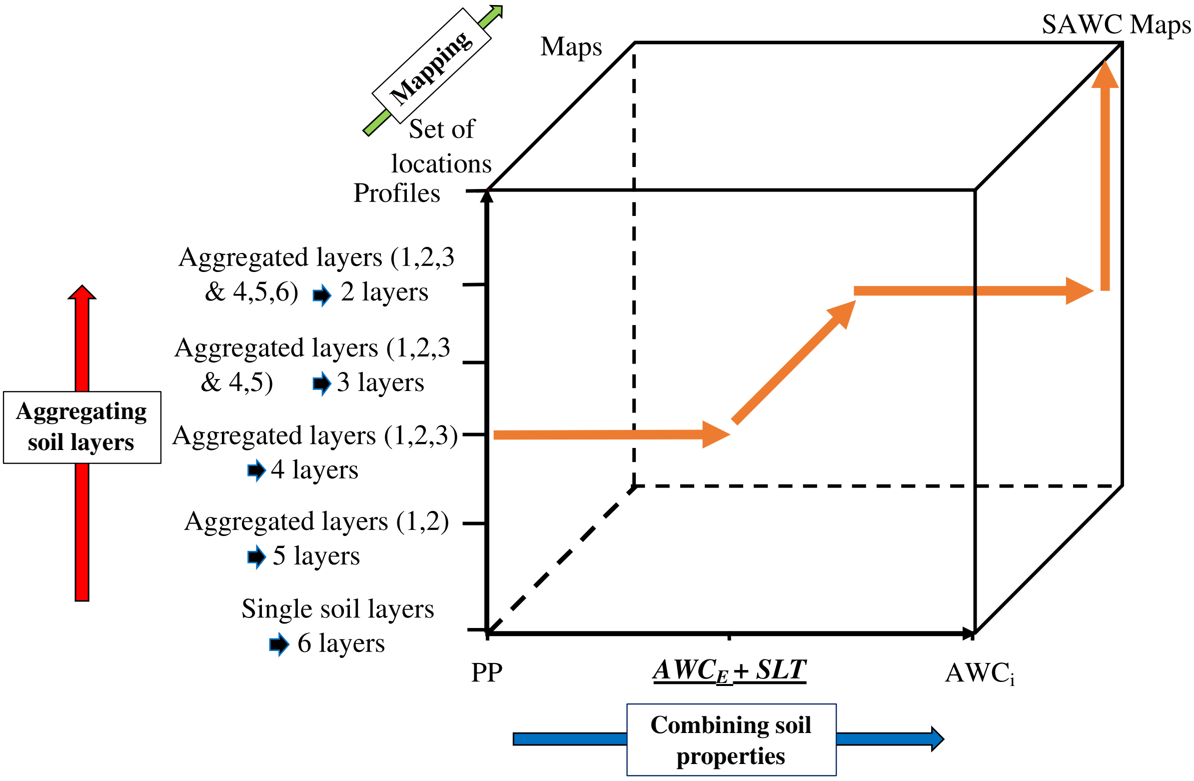


Figure 2. Conceptual diagram of SAWC digital soil mapping with an example of inference trajectory including the new level of soil property combination, elementary available water capacity (AWC_E) and soil layer thickness (modified from Styc and Lagacherie, 2019) (in color)

The rationale of this modification was to separately map two soil properties that exhibited very low correlations, meaning that their variations could result from different landscape drivers that could be imperfectly considered by a single mapping model.

3.4. Uncertainty analysis using error propagation modeling

Following Román Dobarco et al. (2019), the error propagation was modeled using first-order Taylor expansion to calculate the variance of the SAWC predictions. This variance was considered a proxy of the prediction uncertainty of the target variable (Heuvelink et al., 1989). This method relies on the approximation of the estimates obtained for the soil property (i.e., the available water capacity). Let Y be an estimation of a given soil property as follows (Eq. 6):

$$Y = f(z) \tag{6}$$

where f is a continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}$ and z is the vector of the n input variable of f. The approximation of f uses a series centered on the mean values of the n input variables $\mu = [\mu_1, \mu_2, \dots, \mu_n]^T$ (Heuvelink et al., 1989). The variance of Y = f(z) is calculated with the following formula (Eq. 7):

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left(\frac{\delta f(\mu)}{\delta z_i}\right)^2 \sigma z_i^2 + 2 \sum_{\substack{i=1 \\ i \neq k}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} \left(\frac{\delta f(\mu)}{\delta z_i}\right) \left(\frac{\delta f(\mu)}{\delta z_k}\right) \sigma z_i z_k \tag{7}$$

where σ_y^2 is the variance of Y, z_i and z_k are the soil input variables, $\sigma z_i z_k$ is the covariation of the z_i and z_k errors from the i and k variables, σz_i is the standard deviation of z_i and $\frac{\delta f(\mu)}{\delta z_i}$ and $\frac{\delta f(\mu)}{\delta z_k}$ are the partial derivatives of f(z) around μ. σz_i is estimated by the standard deviations of the conditional distributions provided by QRF at each predicted location (Meinshausen, 2006).

Then, the estimate of the variance can be used to compute the limits of the confidence interval. Assuming a normal distribution, the limits of the confidence interval can be computed as follows (Eq. 8):

$$CIL_i = \hat{y}_i \pm 1.645\, \sigma_{\hat{y}_i} \quad (8)$$

where CIL_i is the interval limits of the prediction, $\hat{y}_i$ the mean of the distribution, $\sigma_{\hat{y}}$ the standard deviation and 1.645 is the Student's coefficient for a 90% confidence interval estimation.

Error propagation was performed using the *propagate* R package (Spiess, 2018).

3.5. The experiment

In this study, we considered two sources of uncertainty: i) the mapping error of both SAWC components, i.e., the soil thickness and the AWC_E, and ii) the error of the AWC of every soil layer.

3.5.1. The tested options of error propagation

To evaluate the importance of the correlations of the AWC components errors for the quantification of SAWC uncertainty, four options of error propagation were considered according to whether are considered: i) the error correlations between the predicted properties involved in the determination of AWC at each soil layer, i.e. AWC and soil layer thickness (denoted further SP) ii) the error correlations between the predicted AWC at different soil layers (denoted further SL), iii) both of these correlations (denoted further SP.SL) or iv) none of these correlations (denoted further NONE).

To compute the error correlations, we considered the residuals calculated by the K-fold cross validation (see next section).

Additionally, we derived the SAWC predictions according to three different fixed maximum soil depth, i.e. 200 cm, 100 cm and 60 cm. The rationale was to determine if the predictions for the deepest layers (60-100 cm and 100-200 cm) played a beneficial or nonbeneficial roles in the SAWC predictions.

3.5.2. Evaluation protocol

The performance of the SAWC DSM was evaluated by k-fold cross validation. This evaluation procedure consisted of randomly dividing the data into k subsets. Then, the holdout method was repeated k times such that one of the k subsets was used as the validation set in each repetition, while the other k-1 subsets were merged to form the calibration set. Following this procedure, every data point was included in a calibration set k-1 times. In this study, we selected k = 10; to increase the robustness of the evaluation, the 10-fold cross-validation was iterated 20 times. The k-fold cross-validation was performed using cvTools (Alfons, 2012) that was used to define the folds.

To evaluate the prediction performances, we used classic performance indicators, e.g., the mean square error skill score (SS_{MSE}, Nussbaum et al., 2017), which has the same interpretation as the percentage of variance explained by the model, the root mean square error (RMSE) and the bias.

Furthermore, we evaluated the estimation of the prediction uncertainty using the prediction interval coverage probability (PICP, Eq. 9) (Shrestha and Solomatine, 2006), which was computed as follows:

$$PICP = \frac{count(LPL_i < y_i < UPL_i)}{n} \times 100 \qquad (9)$$

where n is the total number of observations in the validation set, and the numerator counts if the observation y_i fits within the prediction limits prior to estimation by the error propagation method. For a 90% confidence level, which is usually chosen in DSM studies (Arrouays et al., 2014), the uncertainty is optimally predicted when the PICP value is close to 90%.

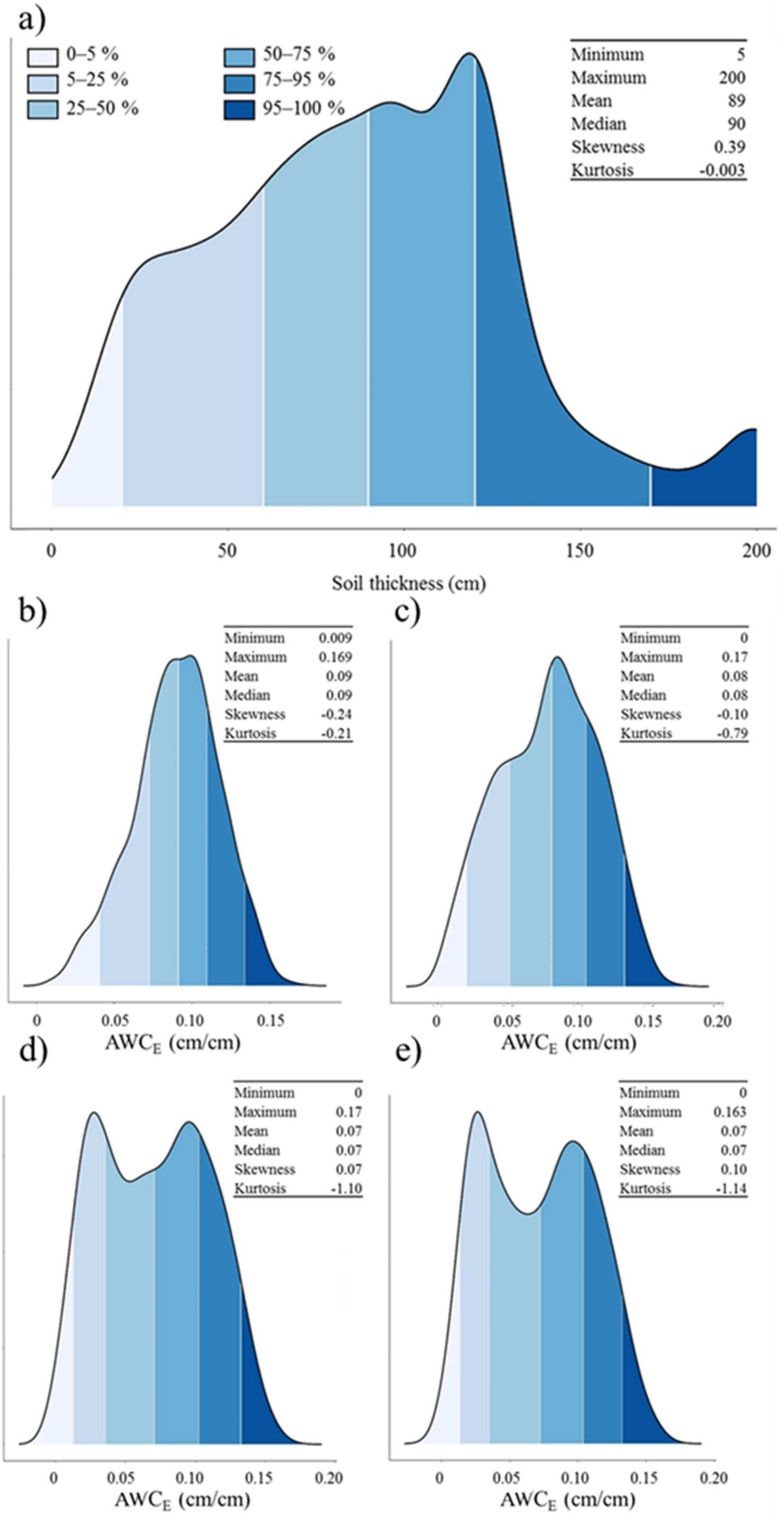
In addition to PICP, we verified that the largest errors were at locations having the largest widths of estimated prediction intervals. For that, the population of validation sites was split into four quartiles of predicted interval widths and four RMSEs were computed separately for each quartile.

4. Results

4.1. Basic statistics

4.1.1. Soil input distributions

In Figure 3, we present the distributions of the soil thickness and the elementary available water capacity across the set of soil profiles that was used as input of the mapping model. The soil thickness ranged from 5 to 200 cm (i.e., the maximum soil observation depth was fixed at 200 cm), with the average ST at 89 cm, which was close to the median value (90 cm). The most common ST in the dataset was 120 cm soil thickness. Then, the distribution dropped dramatically, showing that deep soils were less represented than shallow soils. The shape of the distribution was far from normal, although the skewness and kurtosis tests indicated a normal distribution. The AWC_E for the 0-30 cm and 30-60 cm soil layers were nearly normally distributed, with peaks at 0.09 and 0.10, respectively. The AWC_E for the 60-100 cm and 100-200 cm soil layers differed from the shallowest layers by the shapes of their distributions, which were bimodal (two distribution peaks located at 0.03 cm/cm and at 0.11 cm/cm). While the skewness test indicated that the distributions of the AWC_E values of every soil layer were approximatively symmetric, the excess kurtosis test showed that the distributions were less peaked and presented less extreme values than a normal distribution.



300 AWC$_E$: elementary available water capacity

Figure 3. Distributions of the observed SAWC components at sampling locations ; a) soil thickness, b) elementary available water capacity of the 0-30 cm soil layer, c) elementary available water capacity of the 30-60 cm soil layer, d) elementary available water capacity of the 60-100 cm soil layer and e) elementary available water capacity of the 100-200 cm soil layer (in color)

4.1.2. Correlation of error between the AWC_E and the soil thickness

The correlations between the errors from AWC_E and soil thickness (table 3) were very low whatever the considered soil layer and even non-significant for layers 0-30 cm and 60-100 cm. It may reveal that the mapping models of these soil properties were very different by weighing very differently the used soil covariates.

Table 3. Correlation of the error between the elementary available water capacity (AWC_E) and the soil layer thickness

		Soil properties			
		AWC_E			
	Depth intervals (cm)	0-30	30-60	60-100	100-200
Soil layer thickness	0-30	0.01 ^{ns}	-	-	-
	30-60	-	0.09 **	-	-
	60-100	-	-	0.02 ^{ns}	-
	100-200	-	-	-	-0.13 *

^{ns}: not significant (p-value > 0.05); *: p-value ≤ 0.05; **: p-value ≤ 0.01; ***: p-value ≤ 0.001

4.1.3. Correlation of the errors between soil layers

Table 4 shows the correlation of the AWC errors between soil layers. The AWC errors of the soil layers were correlated, especially for the consecutive soil layers, and the correlations decreased for deeper layers, which may denote large similarities between mapping models of consecutive soil layers. It is worth noting that all error correlations between the soil layers were significant and generally higher than the error correlations between the primary soil properties.

Table 4. Correlation of the available water capacity (AWC) error between the soil layers

Soil property		Depth intervals (cm)		
AWC		0-30	30-60	60-100
	30-60	0.67^{***}	-	-
	60-100	0.42^{***}	0.63^{***}	-
	100-200	0.16^{**}	0.24^{***}	0.36^{***}

AWC: available water capacity

4.2. SAWC component prediction performances

In Table 5, we present the performances of the SAWC component predictions. While the SS_{MSE} values of both the clay and sand contents increased in depth, SS_{MSE} of both the coarse fragment content and soil thickness showed very low performances (e.g., the SS_{MSE} was close to 0%) as well low RMSE and bias values. The performances of the AWC_E indicated a weak SS_{MSE} that ranged from 6 to 14% of explained variance.

Table 5. Prediction performances of soil available water capacity (SAWC) components: Mean values (standard deviations) over 20 iterations of SSMSE (mean square error skill score) RMSE (root mean square error) and bias.

Soil properties	Unit	Depth interval (cm)	SS_{MSE}	RMSE	Bias
Clay	(% mass)	0-30	0.16 (0.007)	10 (0.05)	-2 (0.04)
		30-60	0.20 (0.006)	11 (0.06)	-2 (0.04)
		60-100	0.26 (0.008)	12 (0.07)	-2 (0.09)
		100-200	0.27 (0.011)	12 (0.11)	-2 (0.13)
Sand	(% mass)	0-30	0.29 (0.005)	15 (0.09)	1 (0.08)
		30-60	0.29 (0.006)	18 (0.07)	0 (0.07)
		60-100	0.33 (0.008)	16 (0.11)	0 (0.12)
		100-200	0.32 (0.009)	17 (0.23)	-1 (0.18)
Coarse fragments	(% vol)	0-30	0.07 (0.008)	20 (0.09)	-4 (0.21)
		30-60	0.09 (0.01)	26 (0.15)	-4 (0.12)
		60-100	0.07 (0.013)	29 (0.21)	-3 (0.12)
		100-200	0.03 (0.013)	30 (0.21)	-3 (0.14)
AWC_E	($cm.cm^{-1}$)	0-30	0.12 (0.007)	0.03 (0.0001)	0 (0.0001)
		30-60	0.14 (0.007)	0.03 (0.0001)	0 (0.0001)

		60-100	0.11 (0.008)	0.04 (0.0002)	0 (0.0002)
		100-200	0.06 (0.008)	0.04 (0.0002)	0 (0.0002)
Soil layer thickness	(cm)	0-30	-0.01 (0.005)	3 (0.01)	1 (0.01)
		30-60	-0.04 (0.009)	10 (0.03)	4 (0.02)
		60-100	0 (0.015)	17 (0.11)	4 (0.14)
		100-200	0.01 (0.013)	26 (0.15)	-9 (0.09)

AWC_E: elementary available water capacity

We present in Table 6 the performances of the SAWC predictions obtained by aggregating the predictions of AWC_E and soil thickness of the four aggregated soil layers till the selected maximum rooting depths (60 cm, 100 cm or 200 cm).

SAWC was poorly predicted regardless of the maximum rooting depth considered. The variance explained by the model reached 12-13%. Positive bias values (between 0.51 and 0.97 cm) denoted an overall overestimation of SAWC.

Table 6. Performances of the soil available water capacity (SAWC) predictions : Mean values (standard deviations) over 20 iterations of SS_{MSE} (mean square error skill score) RMSE (root mean square error) and bias.

Maximum rooting depth (cm)	SS_{MSE}	RMSE (cm)	Bias (cm)
60	0.13 (0.008)	1.86 (0.008)	0.51 (0.006)
100	0.12 (0.007)	3.29 (0.013)	0.97 (0.013)
200	0.12 (0.007)	4.2 (0.018)	0.66 (0.017)

4.3. Uncertainty in the SAWC mapping prediction performances

In Table 7, we present the uncertainty evaluation of the predictions averaged with their standard deviations using the PICP. The PICP values ranged from 71% to 91%; the PICP values were closer to optimal (i.e., 90%) when the correlation of the errors between the soil layers was considered regardless of whether the correlation of errors between the soil properties were considered during error propagation. It is worth noting that when the correlation of the errors between the soil layers was not accounted for, the PICP dropped dramatically to values ranging between 71 and 81%, which led to an underestimation of the SAWC uncertainty.

Table 7. Uncertainty prediction evaluation: Mean values (standard deviations) over 20 iterations of PICP (prediction interval coverage probability) with different options for error propagation: considering the error correlation between both soil properties and soil layers (SP.SL), solely the soil layer error correlation (SL), solely the soil property error correlation (SP) or no correlation (NONE).

Maximum rooting depth (cm)	PICP (%)			
	Options for error propagation			
	SP.SL	SL	SP	NONE
60	89 (0.46)	88 (0.40)	81 (0.53)	80 (0.59)
100	88 (0.40)	87 (0.40)	71 (0.59)	71 (0.51)
200	91 (0.28)	91 (0.34)	75 (0.53)	75 (0.48)

Table 8 shows the differences in the prediction performances (RMSE) for the different quartiles in the 90% confidence interval width using the error propagation model for SL (i.e., the one that considered the correlation of the error between soil layers only) for SAWC predicted at 200 cm. The RMSE calculated separately for each quartile tended to increase from the predicted confidence interval with the smallest width to the largest confidence interval (from 3.12 cm to 5.51 cm). Therefore, as expected, the uncertainty predicted by the model was related to the uncertainty observed through the validation protocol. Similar trends were observed for SAWC predictions at 100 and 60 cm.

Table 8. RMSEs for the quartiles of prediction interval widths.

Maximum rooting depth (cm)	Predicted uncertainty (cm)	RMSE (cm)
60	< 5.81	1.67
	5.81 - 6.23	1.86
	6.23 - 6.67	1.95
	> 6.67	1.96
100	<9.45	2.64
	9.45 - 10.39	3.04
	10.39 - 11.23	3.38
	> 11.23	3.96

200	< 11.6	3.12
	11.6 - 13.1	3.77
	13.1 - 15	4.02
	> 15	5.51

4.4. Spatial distribution and uncertainly of the SAWC

4.4.3. Spatial distribution of the SAWC

According to the previously presented results (cf. Table 7), there was no clear difference in performance in the SAWC predictions of 60, 100 or 200 cm, so we chose to present the SAWC map for the maximum soil thickness of 200 cm.

The SAWC map (Figure 5) was mainly divided in two regions of contrasting soil thickness that corresponded to different lithologies and reliefs. The low predicted values of SAWC (shown in red) were predominant in the mountainous crystalline rocks of the Pyrenees and the Massif Central mountains, which were located in the south and northwest of the region, respectively, and on the hard limestone plateaus (the Causses). The high predicted values of the SAWC (in blue color) were located in the hills and plains of the soft marine and fluviatile sediments located near the seaside and in a narrow channel in the west of the region. However, more subtle differences in the predicted SAWC could be observed within the two regions. In the sedimentary area, a gradient was observed from high predicted SAWCs in the alluvium valleys, which had deep soil with low coarse fragment contents, to low predicted SAWCs in the stony soils of the old alluvial terraces (e.g., Nîmes Costières, which is in a red circle in Figure 5) the soils on tertiary sediment (“molasse”) hillsides showed intermediate values. The mountainous crystalline rock areas and the Causses also showed identifiable differences in the predicted SAWCs (dark blue circle in Figure 5) that could be explained by the soil map and DEM derivative covariates (e.g., Multi Resolution Valley Bottom Flatness index, Multi Resolution Ridge Top Flatness index and slope). Similar situations were encountered in the Causses (light blue circles in Figure 5).

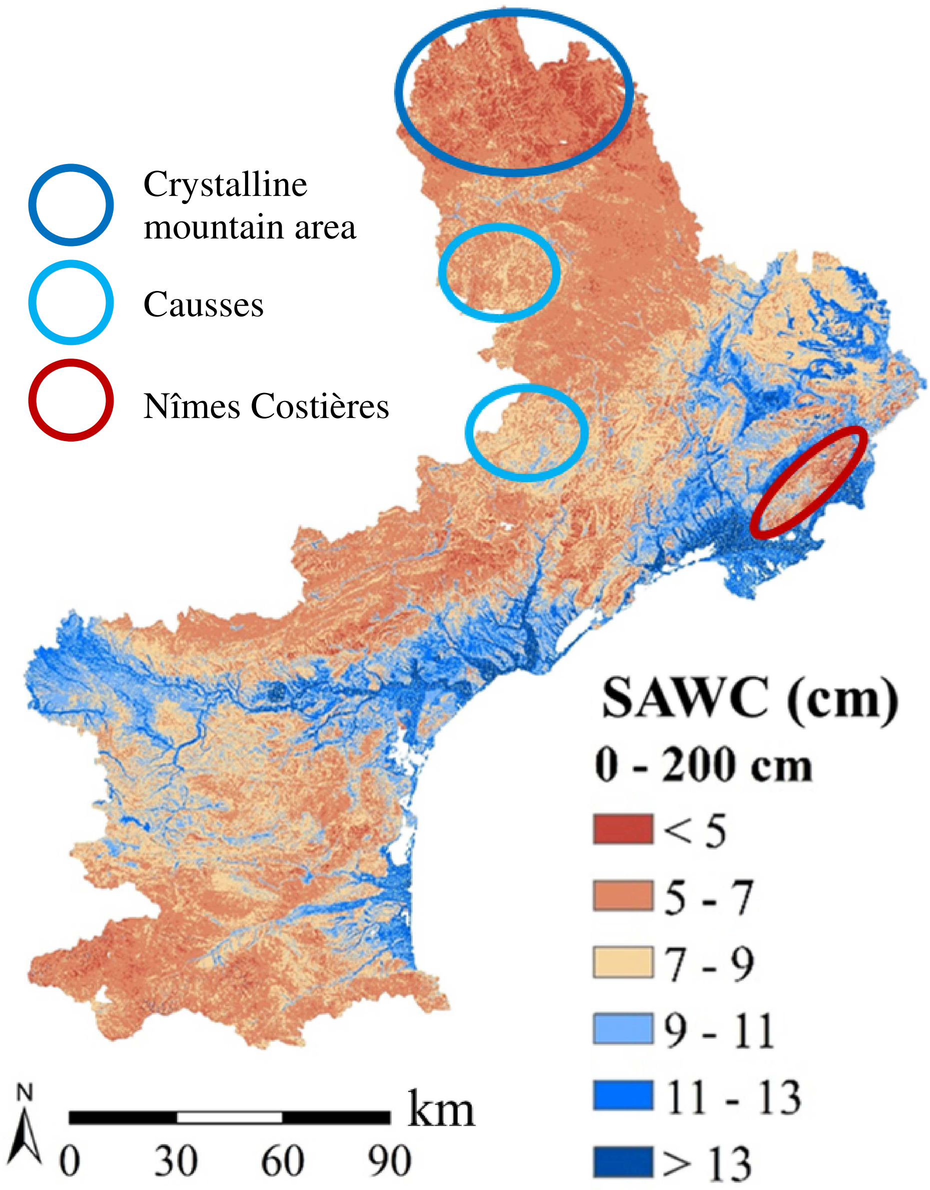


Figure 5. Predicted soil available water capacity (SAWC) map of Languedoc-Roussillon (in color).

4.4.4. Spatial distribution of the SAWC prediction uncertainty

In Figure 6, we provide an uncertainty map estimated as the width of the 90% confident prediction interval represented according to the quartile classes. The high uncertainty predictions were mainly located in the alluvium valley in the littoral region. The moderated uncertainty predictions were located in a large portion of the alluvium valley and in the littoral region, while the low uncertainty predictions occupied the rest of the study area (i.e., the plateaus and mountain regions). It is worth noting that the amount of predicted uncertainty seemed to be mainly related to the values of the predicted SAWC, the biggest errors being related with the predictions of the largest SAWC values.

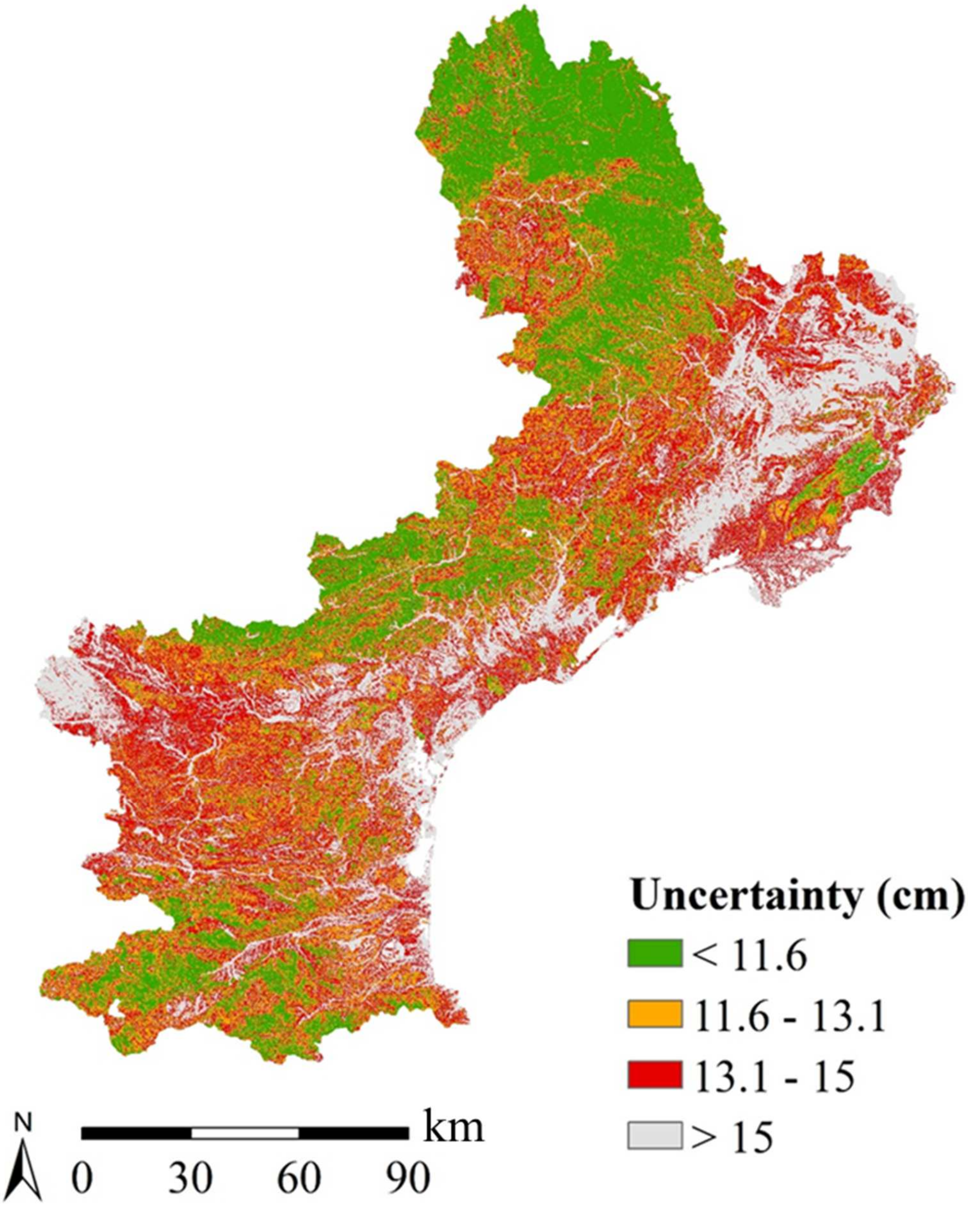


Figure 6. Uncertainty map presented by the classes estimated from the quartiles of the validation distribution (in color)

5. Discussion

5.1. Evaluation protocol

The evaluation protocol that was applied in this study consisted of a 10-fold cross validation approach with 20 repetitions, with the reference data being the soil profiles at punctual sites with the observed properties (textural fractions, coarse fragments and soil thickness) used to produce a local estimation of the SAWC using the pedotransfer functions. This protocol ensured both, an evaluation of the SAWC predictions from independent samples and a comprehensive consideration of the mapping errors of all SAWC components. To our knowledge, this is the first time that these two conditions were fulfilled in an evaluation protocol dealing with the SAWC, which makes the comparisons with previous papers dealing with SAWC (e.g. Poggio et al, 2010; Hong et al, 2013; Leenaars et al, 2018) difficult. However, our evaluation protocol had two main limitations. First, the evaluation sites were not characterized by the real SAWC measurements, and the fine earth water retention was estimated with the pedotransfer functions. This did not allow us to account for the PTF errors, as Román Dobarco et al. (2019) did in their study. However, these authors showed that PTF errors played a minor role in comparison with the mapping errors of the SAWC components. This should be even more true with the addition of the ST mapping errors in the evaluation protocol in comparison with the protocol of Román Dobarco et al (2019). Furthermore, the introduction of the PTF error in our error propagation models following the technique proposed by Román Dobarco et al. (2019) did not modify the ex-ante uncertainty evaluation of the SAWC paper (result not shown in this paper). However, a full evaluation of the SAWC mapping would be preferable, which means investing in costly field and laboratory measurements or finding accurate and inexpensive proxies for the SAWC (Coulouma et al, 2020).

A second limitation of our evaluation protocol was that SAWC mapping was evaluated at the site level, which was not representative of the decision-making units of the end-user and represents the worst case scenario for assessing the soil property prediction quality. Ideally, the evaluation should be performed for areal units (Bishop et al, 2015), which would produce more realistic results that would be in accordance with the visual evaluation of the map (see further). However, evaluating the SAWC from the areal units would require data collection that cannot be reasonably envisaged. Consistency checking involving data available at larger spatial support and closely related with AWC could be an alternative (Vanderlinden, 2005)

5.2. Error propagation model

Following Román Dobarco et al. (2019), an error propagation model using a first-order Taylor expansion was developed for the ex-ante estimation of SAWC mapping uncertainty. This model was, to some extent, more complete than the one developed by Román Dobarco et al. (2019) in that it considered the error in the soil thickness maps and the correlations between the error in the SAWC component maps that were not considered in Román Dobarco et al. (2019). This model was obtained by selecting an inference trajectory that included separate mapping of the soil layer thicknesses, which allowed easy application of the first-order Taylor expansion. The results revealed that the consideration of the error correlations had an impact on the final result if they reached a given level, which was the case for the error correlations between the soil layer maps (Table 4). The results showed that the ex-ante estimation of uncertainty was only slightly biased (differences with nominal values of 90% less than 1 for two out of three SAWC maps), which corresponded to much smaller uncertainty estimation biases than those obtained by Román Dobarco et al. (2019). We also verified that the RMSE obtained from the validation protocol was closely related to the predicted uncertainty (Figure 4), which, to our knowledge, has not yet been verified. We note, however, that the error propagation model built in this study did not consider the PTF errors. This error could be easily added by following the procedure proposed by Román Dobarco et al. (2019).

5.3. General performances of the SAWC predictions

Although the SAWC map of the Languedoc-Roussillon region exhibited expected and pedologically sound soil patterns, poor results were obtained from the evaluation protocol (Table 6). This was likely related to the difficulty of mapping the two most critical components of the SAWC, namely, the coarse fragments and the soil layer thicknesses (Table 5). As observed by Vaysse et al. (2015), the soil thickness and coarse fragments were characterized by a pre-eminence of short-scale variations that could not be captured by a DSM model using a so sparse soil spatial sampling according to the soil layer depths (Table 1). A denser spatial sampling is therefore necessary in this situation. Furthermore, the coarse fragment data were obtained from visual observations of the soil profile, which carry a greater uncertainty than the ones of other soil properties that are measured in a laboratory. More accurate field protocols for measuring the proportion of coarse fragments (Algayer et al, 2019) are required to improve this situation.

It is also important to notice that biases were important components of the SAWC prediction errors (between 15 and 30% as shown on table 6), which generated an overall SAWC overestimation. This overestimation can be related with the important positive biases observed for the predictions of the thicknesses of the three first soil layers (table 5). Such biases should be caused by the difficulties of the Random Forest algorithm to deal with the important subset of locations having null soil layer thicknesses. Dealing with zero-inflated input datasets of regression models is a well-known problem in ecology (Martin et al, 2005). Specific regression approaches adapted to zero-inflated datasets (Savage et al, 2015) should be applied to mitigate this problem.

Conversely, the last soil layer exhibited a negative bias (table 5) that can be related with the unbalanced distribution of soil thicknesses in the set of sampling locations (figure 3). Indeed, a very small proportion of deep soils were sampled because of the low maximum observation depths (only 15% of the soil profiles had maximum observation depths greater than 120 cm).

This too small proportion generated the underestimation of the deepest soil layers thicknesses since, the random forest algorithm is known to behave like an interpolator and to smooth the outliers (Song, 2015). A more balanced sampling of the ST across the study region is therefore necessary.

6. Conclusions

- We developed a DSM model that mapped the SAWC values and provided an ex ante local estimation of the prediction uncertainty. For the first time, this uncertainty model took into account all SAWC component mapping errors for all soil layers.
- The results showed weak performances of the SAWC predictions, although the final map exhibited pedologically sound spatial patterns of predicted SAWC. This paradoxical result could be caused by the inadequate spatial support at which the evaluations were conducted (punctual one).

7. Acknowledgments

This research was granted by the French National Research Institute for Agriculture, food and environment (INRAE), the French Research and Technology Agency (ANRT) and BRL Exploitation. Philippe Lagacherie is a collaborator of the GLADSOILMAP Consortium supported by the LE STUDIUM Loire Valley Institute for Advanced Research Studies

8. References

Alfons, A. cvTools: Cross-Validation Tools for Regression Models. R Package. 2012. Available online: https://cran.rproject.org/web/packages/cvTools/cvTools.pdf. (accessed on 11 May 2012).

Algayer, B., Lagacherie, P., Lemaire, J., 2020. Adapting the available water capacity indicator to forest soils : An example from the Haut-Languedoc (France). Geoderma 357, 113962.

Al Majou, H., Bruand, A., Duval, O. 2008a. Use of in situ volumetric water content at field capacity to improve prediction of soil water retention properties. Can. J. Soil Sci., 88(4), 533–541.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014. The GlobalSoilMap project specifications, in: GlobalSoilMap: Basis of the Global Spatial Soil Information System - Proceedings of the 1st GlobalSoilMap Conference.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modeling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45.

Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241–242, 238 – 249.

Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

Bruand, A., Pérez Fernández, P., Duval, O., 2003. Use of class pedotransfer functions based on texture and bulk density of clods to generate water retention curves. Soil Use Manage. 19, 232–242.

Coulouma, G., Prevot, L., & Lagacherie, P. 2020. Carbon isotope discrimination as a surrogate for soil available water capacity in rainfed areas: A study in the Languedoc vineyard plain. Geoderma, 362, 114-121.

Cousin, I.,Nicoullaud, B., Coutadeur, C. 2003. Influence of rock fragments on the water retention and water percolation in a calcareous soil. Catena, 53, pp. 97-114

Dobarco, M.R., Bourennane, H., Arrouays, D., Saby, N.P., Cousin, I. and Martin, M.P., 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma, 344, pp.14-30.

Dominati,E., Mackay, A., Green, S., Patterson, M. 2014.A soil change-based methodology for the quantification and valuation of ecosystem services from agro-ecosystems: a case study of pastoral agriculture in New Zealand. Ecological Economics. Econ., 100, pp. 119-129

Heuvelink, G.B., 2014. Uncertainty quantification of GlobalSoilMap products, GlobalSoilMap: basis of the global spatial soil information system. Proc. 1st Glob.Soil.Map. Conf. 335-340.

Heuvelink, G.B.M.; Burrough, P.A.; Stein, A. 1989. Propagation of errors in spatial modelling with GIS Int. J. Geogr. Inf. Syst., 3, 303–322.

Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., Lee, K., 2013. Predicting and mapping soil available water capacity in Korea. PeerJ, 1:e71

Leenaars, J.G.B., Claessens, L., Heuvelink, G.B.M., Hengl, T., Ruiperez González, M., van Bussel, L.G.J., Guilpart, N., Yang, H., 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.

Martin, T.G., Wintle, A., Rhodes, J.R., Field, A., Low-choy, S.J., 2005. Zero tolerance ecology : improving ecological inference by modelling the source of zero observations. Ecol. Lett. 1235–1246.

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

Meinshausen, N. 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L.; Schaepman, M.; Papritz, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 2017, 4, 1–22.

Poggio, L., Gimona, A., Brown, I., Castellazzi, M., 2010. Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. Geoderma, 160(2), 175–188.

Román Dobarco, M., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma, 336, 81–95.

Savage, S.L., Lawrence, R.L., Squires, J.R., 2015. Predicting relative species composition within mixed conifer forest pixels using zero-in fl ated models and Landsat imagery. Remote Sens. Environ. 171, 326–336.

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 19 (2), 225–235

Song, J. (2015). Bias correction for Random Forest in regression using residual rotation. Journal of the Korean Statistical Society, 44, 321-326.

Spiess, A-N., 2018. Propagate: Propagation of Uncertainty. R package version 1.0-6. Available on : https://CRAN.R-project.org/package-propagate

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348.

Styc, Q.; Lagacherie, P. What is the Best Inference Trajectory for Mapping Soil Functions: An Example of Mapping Soil Available Water Capacity over Languedoc Roussillon (France). Soil Syst. 2019, 3, 34.

Ugbaje, S.U., Reuter, H.I., 2013. Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. Vadose Zone J., 12(4).

Vanderlinden, K., Giráldez, J. V., Van Meirvenne, M., 2005. Soil water-holding capacity assessment in terms of the average annual water balance in southern Spain. Vadose Zone J., 4, 317–328.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, 291, 55–64.

Vaysse, K.; Lagacherie, P. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg. 2015, 4, 20–30

Veihmeyer, F.J., Hendrickson, A.H. 1927. The relation of soil moisture to cultivation and plant growth. Soil Sci. 3, 498–513.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J. 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of Hydrology (Amsterdam), 251, pp. 123-150.

Wright, M.; Ziegler, A. 2017. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw., 77, 1–17.

INRAE