Plant Soil (2023) 492:501–517
https://doi.org/10.1007/s11104-023-06198-x

RESEARCH ARTICLE

The power of integrating proximal and high-resolution remote sensing for mapping SOC stocks in agricultural peatlands

Wanderson de Sousa Mendes · **Michael Sommer** · **Sylvia Koszinski** · **Marc Wehrhan**

Received: 25 May 2023 / Accepted: 27 July 2023 / Published online: 3 August 2023

Abstract

Background and aims Soil electrical conductivity (ECa) data derived from electromagnetic induction (EMI) is valuable for estimating peat thickness and soil organic carbon stocks (SOC_{stocks}). However, generating ECa maps locally using geostatistics limits the coverage area. This study explores the use of digital soil mapping (DSM) with random forest (RF) and universal kriging (UK) to create high-resolution ECa maps from field survey EMI data. The objective is to enhance the predictive accuracy of SOC_{stocks} models in peatlands by incorporating these ECa maps as environmental variables.

Methods Three scenarios were evaluated, combining different environmental variables and modelling techniques for ECa mapping. Scenario 1 used spectral indices from RapidEye satellite data and RF. Scenario 2 included spectral indices and terrain derivatives from LiDAR, with RF. Scenario 3 integrated spectral indices, terrain derivatives from LiDAR, and UK. Afterwards, we evaluated the effectiveness of adding ECa maps as environmental variables for SOC_{stocks} mapping. Finally, we incorporated ECa maps from scenario 2 and RF in three ways: (a) scenario 2 variables only, (b) ECa_2 with scenario 2 variables, and (c) ECa_3 with scenario 2 variables.

Results Scenarios 2 (ECa_2) and 3 (ECa_3) outperformed scenario 1 (ECa_1). The inclusion of ECa maps significantly improved the accuracy of SOC_{stocks} models.

Conclusion Our study demonstrates that DSM, combined with RF and UK techniques, enables the generation of high-resolution ECa maps from field surveys in peatlands. Incorporating these ECa maps as environmental variables enhances the accuracy of SOC_{stocks} mapping, providing valuable insights for peatland management and carbon stock estimation.

Keywords Proximal sensing · Remote sensing · Peat · Climate change · Greenhouse effects · Apparent electrical conductivity

Responsible Editor: Zucong Cai.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s11104-023-06198-x.

W. de Sousa Mendes (✉) · M. Sommer · S. Koszinski · M. Wehrhan
Landscape Pedology Working Group, Leibniz Centre for Agricultural Landscape Research (ZALF), Research Area 1 “Landscape Functioning”, 15374 Müncheberg, Germany
e-mail: wanderson.mendes@zalf.de

M. Sommer
Institute of Geography and Environmental Science, University of Potsdam, 14476 Potsdam, Germany

Introduction

Peatlands are significant global ecosystems, encompassing approximately 3% of the Earth’s land surface.

These ecosystems are crucial carbon reservoirs, storing a substantial amount of soil organic carbon, estimated to be between 450 and 650 Pg (Yu et al. 2010; Loisel et al. 2017; Jackson et al. 2017). Due to their substantial carbon stocks, peatlands play a vital role in the global carbon cycle and are among the largest terrestrial carbon pools. Peatlands have been used for agriculture (e.g., grazing) for over a millennium, increasing heavy drainage, currently raising international concerns. In Germany, for instance, most peatlands no longer work as carbon sinks due to agricultural activities (Tiemeyer et al. 2016, 2020).

Peatlands differ from other ecosystems because of waterlogged conditions, very peculiar vegetation composition (e.g., sedge, reed, and cattail plants), and surface peat layer (Gorham 1991; Lees et al. 2018). Although there are several types of peatlands, they all have the common feature to accumulate a significant volume of organic carbon over a long time period (Minasny et al. 2019). Therefore, quantifying the soil organic carbon stocks (SOC_{stocks}) in peatlands is crucial. However, mapping the distribution of peatlands and their SOC_{stocks} is a tough task (Koszinski et al. 2015; Minasny et al. 2019). Topography, restricted sampling depths, small fragments of peats worldwide, nutrient availability, land use history, and carbon biogeochemical dynamics are the foremost factors that hinder estimation of SOC_{stocks}.

Proximal sensing in geophysical and soil studies through electromagnetic induction (EMI) instruments, which retrieve apparent electrical conductivity (ECa), have been a useful tool to estimate peat thickness and SOC_{stocks} (Saey et al. 2012; Koszinski et al. 2015; Huang et al. 2016). Moreover, ECa information has proven to be highly correlated with peat properties, such as water content, dissolved ions content, and decomposition stages (Comas and Slater 2004; Walter et al. 2015), which are essential data to estimate SOC_{stocks}. ECa has the potential to improve the SOC_{stocks} predictions in soils, mainly peatlands, corroborated by some studies (Comas and Slater 2004; Kettridge et al. 2008) that presented the geophysical methods to quantify peat thickness and SOC_{stocks}.

The main challenge is to provide the most efficient spatial extrapolation of ECa and use it as an environmental variable to improve the predictive power of SOC_{stocks} models. Although there have been several studies on ECa mapping using geostatistical methods (Sun et al. 2012; Saey et al. 2012; Altdorff et al. 2016; Zhang et al. 2020), the use of machine learning methods, such as random forest, for this purpose is relatively scarce in the soil science literature. As evidence of this, we identified only one study in the literature (Taghizadeh-Mehrjardi et al. 2014) that evaluated the creation of digital maps of ECa using regression kriging, a hybrid geostatistical method that combines elements of regression analysis (e.g., machine learning) and kriging to improve spatial predictions of the variable of interest (Keskin and Grunwald 2018).

Digital soil mapping (DSM) uses environmental variables that explain the response variable behaviour in the landscape to map it. To date, few studies have investigated the use of DSM framework to map ECa data from laboratory and EMI. For example, Yang et al. (2019) compared multiple linear regression, geographically weighted regression, mixed geographically weighted regression and the regression kriging of residuals of three algorithms to predict soil electrical conductivity data from laboratory. Wu et al. (2018) also tested three machine learning methods, such as support vector machine, multiple linear regression, and random forest to predict ECa data from EMI. Although those authors utilised ECa data from EMI, they did not collect data following the conventional field sampling design required for geostatistical applications, nor did they incorporate the final predicted ECa maps as environmental variables for modelling other soil properties. Moreover, their analysis was limited to up to 30 spatially distributed samples within their study site.

DSM framework also requires a good source of environmental variables to better explain spatially the response of the variable behaviour in the landscape. Remote sensing (RS) data from satellites is the main provider of environmental variables for DSM (McBratney et al. 2003; Hartemink et al. 2020; Thompson et al. 2020). Satellite images from Landsat, Sentinel, RapidEye, and other missions provide the soil multispectral response and vegetative targets on the Earth surface allowing to calculate their representative indices or use their individual bands separately in different spatial scales (Grinand et al. 2017).

Spectral indices based on satellite imagery, such as brightness index, normalised difference vegetation index, red-edge normalised vegetation index, enhanced vegetation index, and redness index proved to be more effective than use the individual bands separately to predict SOC (Lamichhane et al. 2019), because vegetation is positively correlated to SOC. The authors also highlighted the importance of other RS data, such as digital elevation models (DEM) and their terrain derivatives, because these derivatives can retrieve potential depositional, erosional, moist, and dry areas, which are highly correlated to SOC. Therefore, RS imagery, DEM, and terrain derivatives are recommended to be applied as environmental variables to map SOC.

Spatial resolution of RS data is one of the main factors to increase modelling performance of SOC. Wiesmeier et al. (2019) conducted a comprehensive literature review on different factors that may affect SOC quantification and spatialisation at different scales. High resolution RS data can provide more detailed and effective information of SOC dynamics in the soil surface and subsurface. Consequently, the quality of environmental variables influences the accuracy of SOC prediction (Miller et al. 2016). For example, Forkuor et al. (2017), investigated the use of high resolution data to map SOC spatial distribution and stated that detailed RS data plays an important role in SOC modelling.

There are no reports in the literature of the use of ECa data from EMI with high resolution RS data and machine learning method (e.g., random forest) or universal kriging for DSM. This means that the geostatistical approach at local scale is still the usual procedure to produce high-resolution ECa maps from field survey data. Unlike other geostatiscal methods, the universal kriging method allows kriging in the presence of strong trends in the sample data using available secondary information (i.e., environmental variables) at all prediction locations (Trangmar et al. 1986). Nevertheless, the model parameter fitting is linear due to the strict assumptions of the universal kriging (UK) model. The machine learning methods, such as random forest (RF), work differently from the geostatiscal approach because they do not require strict model assumptions. In this sense, RF can handle non-linear and linear relationships between the response and environmental variables, which allows classifying it as a nonparametric method. RF is also a well-established machine learning algorithm in digital soil mapping (Khaledian and Miller 2020; Wadoux et al. 2020).

Based on these considerations, we hypothesised whether the use of the machine learning method and UK through DSM framework can be viable to produce ECa maps from field surveys and then use ECa maps as an environmental variable to improve the predictive power of SOC_{stocks} models in peatlands. To test the hypotheses, we evaluated three possible scenarios combining different sets of environmental variables with the modelling technique to map ECa. Defined as: Scenario 1, only spectral indices calculated from RapidEye satellite collection and random forest (final map named as ECa_1); Scenario 2, spectral indices calculated from RapidEye satellite collection combined with terrain derivatives from LiDAR sensor and random forest (final map named as ECa_2); and Scenario 3 spectral indices calculated from RapidEye satellite collection combined with terrain derivatives from LiDAR sensor and UK (final map named as ECa_3).

Afterwards, we assessed the efficacy of adding ECa maps as environmental variables to produce SOC_{stocks} maps. Therefore, we incorporated the ECa maps produced (i.e., ECa_2 and ECa_3) with the environmental variables from Scenario 2 and random forest for DSM of SOC_{stocks} as follows: (a) only the environmental variables from Scenario 2, (b) ECa_2 with the environmental variables from Scenario 2, and (c) ECa_3 with the environmental variables from Scenario 2.

Materials and methods

Study site and main landscape features

The study site is part of the Havelländisches Luch (NE Germany) and covers 26.54 km^2 in the municipalities of Wiesenaue, Paulinenaue, and Fehrbellin (Fig. 1). It is also in the major geographical region of Northern Lowland, formed during the Pleistocene era resulting from sub-, peri-, glacio-fluvial and post-glacial geomorphological processes. Topographically, the study site is characterised by flat relief (Fig. 1a), which contributes to the formation of groundwater dependent wetlands or ecosystems. Peatland (water-rise mire) is the main type of wetland in that region (Mueller et al. 2007) and the average groundwater levels are near the surface up to 0.3 m. This area is highly heterogeneous due its parent material (e.g., fluvial sediments, outwash, sand bars, dunes, and alluvial fans) and land management, mainly the drainage history. The mean annual precipitation and

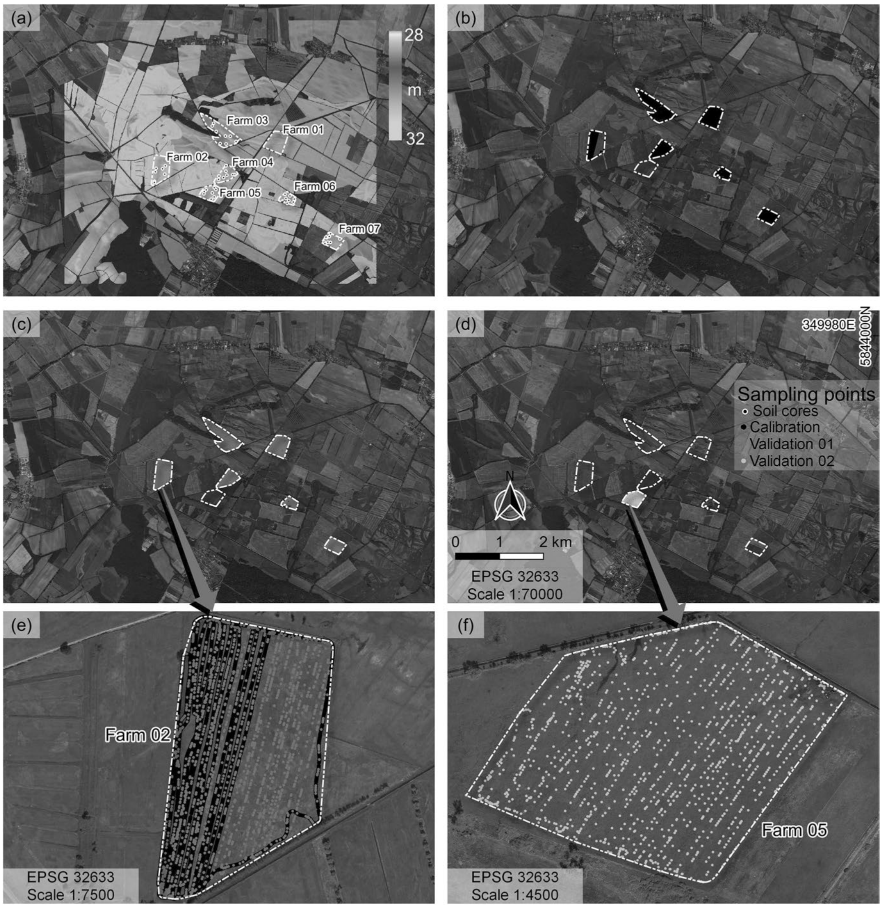

Fig. 1 Study area (©Google Maps 2019) displaying the Digital Elevation Model from LiDAR and collected soil cores (**a**), ECa calibration data representing 80% of the total samples (**b**), ECa validation data representing 20% of the total samples named validation 01 (**c**), and additional validation data outside the total data named validation 02 (**d**) from 2010. Validation 01 involves cross-validation (**e**), while Validation 02 employs hold-out validation (**f**), as exemplified by zooming in on Farm 02 and 05

air temperature are 550 mm (interannual variation 340–950 mm) and 9.5 °C, respectively. The most common soil groups are Arenosols, Gleysols, and Histosols (IUSS Working Group WRB 2015) used for arable land, meadows, and pastures.

Soil and apparent electrical conductivity data

Data on total SOC_{stocks} were acquired from 49 soil cores (Fig. 1a) and stratified random sampling was the selection procedure of those points, as described in Koszinski et al. (2015). The soil samples were collected using a hydraulic probe (2 m long and 10 cm in diameter), air-dried, ground, and sieved (< 2 mm mesh). Only the fraction < 2 mm was used for further analyses. An aliquot of the air-dried < 2 mm fraction was oven-dried at 105 °C (24 h) to obtain the water contents of the air-dried samples. The total C content was determined using the elemental analysis (CNS analyser TruSpec, LECO Ltd., Mönchengladbach, Germany) as CO_2 through infrared detection after dry combustion at 1250 °C (DIN ISO10694 1996). The gas-chromatographic analysis of CO_2 evolution (Carmograph by Woesthoff, Scheibler-method) was used to determine carbonate C after an application of phosphoric acid. The subtraction between total and carbonate C results in the soil organic carbon content (SOC_c). All data are presented on an oven-dry basis.

Steel rings of 100 cm^3 were used to collect soil samples from 103 out of 282 horizons (49 soil profiles). Afterwards, bulk density (D_b) was determined using the thermogravimetric desiccation at 105 °C (DIN EN ISO 11272 2014) in the lab. The remained 179 missing values of D_b were calculated using a pedotransfer function (e.g., relationship between measured D_b and SOC_c), as described in Koszinski et al. (2015), and not collected because they were inaccessible (groundwater levels or disturbed by the sampling). Thus, SOC_{stocks} for each soil core was calculated to a 1 m depth following the methodology and formula described in Schlichting et al. (1995):

$$SOC_d = \sum_{i=1}^{N} SOC_c \times t \times D_b \left[(100 - Gr)/100 \right]$$

Where: SOC_d is the soil organic carbon density (SOC_{stocks}, kg m^{-2}), SOC_c is the soil organic carbon content (wt%), t is thickness (cm), D_d is bulk density (Mg m^{-3}), and Gr is the mass percentage of the fraction > 2 mm (gravel, wood fragments etc.) of a specific horizon.

Apparent electrical conductivity (ECa) was measured using an electromagnetic induction sensor with the minimal contact EM38DD device (Geonics Ltd., ON, Canada) in April (farms 01, 02, 03, 06, and 07), and July – August (farms 04, and 05) of 2010. The device retrieves ECa in two dipole sets, horizontal (ECa_h, 0.75 m), and vertical (ECa_v, 1.5 m) (Abdu et al. 2007). In our study, we selected ECa_v, described as ECa throughout this article, because it has the same depth information as calculated SOC_{stocks}. The effect of soil temperature was measured at a 0.1 m depth to calibrate the device and standardised at 25 °C (Brevik et al. 2004). ECa data were collected in eight different fields with tracking lanes of 10–15 m apart and within lanes of 1–3 m (Koszinski et al. 2015). The total number of samples was 59,759. ECa data from 2010 was used to calibrate (Fig. 1b) and validate (Fig. 1c and d) the models.

Environmental variables

Eight high-resolution environmental variables were generated. Four variables derived from an airborne high-resolution Light Detection and Ranging (LiDAR) and four from RapidEye satellite images to model ECa (Table 1). The Digital Elevation Model (DEM) was retrieved from LiDAR in December 2008. The basic outputs to generate the DEM were four points per square meter at a 50,000 Hz pulse frequency, and vertical accuracy of 3 cm in the *z-axis*. DEM was generated using the scattered points by aggregating on a 1 × 1 m grid and resampled to 5 m by the nearest neighbour method. Three terrain derivatives were calculated through QGIS (QGIS Development Team 2020) using the DEM. The first was the Topographic Position Index (TPI) in which positive values characterise locations higher than the mean of their surroundings, while negative values stand for locations lower than their surroundings. TPI near or equal zero are either constant slope areas or

Table 1 Environmental variables retrieved from Light Detection and Ranging (LiDAR) and RapidEye satellite

Environmental variables	Unit	Spatial resolution (m)	Type of variable	Characteristics	Reference
Digital Elevation Model (DEM)	m	*1	Continuous	Relief	-
Topographic Position Index (TPI)	-	*1	Continuous	Terrain position	(Wilson and Gallant 2000; Weiss 2001)
Slope	%	*1	Continuous	Relief inclination	-
Topographic Wetness Index (TWI)	-	*1	Continuous	Soil water content	-
Normalised Difference Vegetation Index (NDVI)	-	5	Continuous	Healthy and green vegetation	(Rouse et al. 1973)
Red-Edge Normalised Difference Vegetation Index (RENDVI)	-	5	Continuous	Crop stress detection – Chlorophyll content and Nitrogen	(Gitelson and Merzlyak 1994; Sims and Gamon 2002)
Enhanced Vegetation Index (EVI)	-	5	Continuous	Higher Leaf Area Index	(Huete 2004)
Brightness	-	5	Continuous	Average reflectance magnitude	(Agbu et al. 1990)

* Data retrieved from LiDAR and resampled to 5-m spatial resolution

flat areas. TPI is an appropriate relief environmental variable, as it is an intrinsically scale-dependent phenomenon. The other derivatives were the slope and topographic wetness index (TWI), which represent the steepness and possible accumulated water at each pixel of the DEM, respectively. The derivatives were selected because of their high correlation with ECa and SOC_{stocks} (Koszinski et al. 2015).

The RapidEye orbit is composed of five satellites with identical sensors launched in 2008. Each satellite has five bands at a spatial resolution of 6.5 m resampled to 5 m during pre-processing: blue (440–510 nm), green (520–590 nm), red (630–685 nm), red-edge (690–730 nm), and near-infrared (NIR, 760–850 nm). The RapidEye data used in this study were retrieved from the RapidEye Sciences Archives (RESA) for 3rd of June, 27th of June, 3rd of July, 16th of July, 21st of August, and 4th of October 2010, consisting of a 3 A level ortho products (cloud cover $<5\%$). This means that the satellite data were radiometrically and geometrically corrected. Finally, we used the FLAASH algorithm (ENVI V. 4.4.) for atmospheric correction. Scattered clouds and respective shadows were removed manually. Moreover, we removed artefacts, such as roads, forest, and urban areas, using vector layers provided by the "Amtliches Topographisch-Kartographisches Informationssystem" (ATKIS). The normalised difference vegetation index (NDVI), red-edge normalised vegetation index (RENDVI), enhanced vegetation index (EVI), and brightness index were calculated from the RapidEye data for each of the specified dates (Table 2).

Modelling

In this study, we evaluated random forest (RF) and universal kriging (UK) algorithms that could satisfactorily deal with ECa and SOC_{stocks} dataset. RF can handle non-linear and linear relationships between response and predictor variables, which classifies RF as a nonparametric method. RF is also a well-established algorithm in machine learning for DSM (Khaledian and Miller 2020; Wadoux et al. 2020). RF trains bootstrap samples of the data using a large number of individual tree models and RF main tuning parameters are the number of variables available for selection in each split (mtry) as well as the number of trees (ntree) (Breiman 2001; Houborg and McCabe 2018). Unlike other geostatiscal methods, UK allows kriging in the presence of strong trends in the sample data using available secondary information (i.e., environmental variables) thus UK can effectively derive prediction uncertainties through kriging variance handling both the variogram modelling of the residuals and regression model together. In our study, the predictions using UK were made on a regular 5 m x 5 m grid. The UK prediction variance simplistically become (Trangmar et al. 1986; Christensen 2011):

$$Z(x) = \beta_0 + \sum_{i=1}^{p} \beta_i f_i(x) + e(x) \quad (5)$$

Where: $Z(x)$ represents the predicted value at location x, β_0 and β_i stand for unknown regression coefficients, $f_i(x)$ denotes the environmental variables, and $e(x)$ indicates a normally distributed residual with zero-mean and constant variance c(0). The residual e may exhibit spatial autocorrelation, which is quantified through a variogram.

Therefore, three possible scenarios were generated by employing distinct sets of environmental variables along with a modelling technique to map ECa. Scenario 1 involves solely spectral indices calculated from the RapidEye satellite collection and random forest; Scenario 2 incorporates spectral indices from the RapidEye satellite collection in combination with terrain derivatives from LiDAR sensor data and random forest; and Scenario 3 comprises spectral indices from the RapidEye satellite collection along with terrain derivatives from LiDAR sensor data and universal kriging. We applied RF to model ECa

Table 2 Vegetation indices selected for this study site

Indices	Formula	
NDVI	$(NIR - red)/(NIR + red)$	[1]
RENDVI	$(NIR - red\ edge)/(NIR + red\ edge)$	[2]
EVI	$2.5 \times \left[(NIR - red)/(NIR + 6 \times red - 7.5 \times blue + 1)\right]$	[3]
Brightness	$(blue + green + red + rededge + NIR)/5$	[4]

and SOC$_{\text{stocks}}$ data, while UK was applied to model ECa data only. For RF models, the calibration data for model fitting were set by using a ten-fold cross-validation method, executed ten times, to avoid the effects of spatial autocorrelation between data points (Wadoux et al. 2021) through the "caret" R package (Kuhn 2008). The RF tuning hyperparameters were all set to their default value by taking into account the lowest root mean square error (RMSE) and highest model efficiency coefficient (MEC) in the "caret" R package. The UK model was performed using the "automap" R package (Gianola et al. 2011). The model and semivariogram are described in Fig. S1.

Evaluation metrics

An exploratory analysis was performed by calculating the Pearson's correlation coefficient (r) between the response and environmental variables. The interpretation of r was based on the following categorisation: low or weak correlation for $-0.35 \le r \le 0.35$, modest or moderate correlation for $0.36 \le r \le 0.67$ and $-0.67 \le r \le$ -0.36, and strong or high correlation for $0.68 \le r \le 1$ and $-1 \le r \le$ -0.68 (Taylor 1990).

The modelling calibration and validation for ECa were performed using the 58,481 observations from 2010. The conditioned Latin Hypercube Sampling (cLHS) was used to split the data in calibration (80%, 46,784 observations)(Fig. 1b) and validation (20%, 11,697 observations)(Fig. 1c and e) based on remote sensing data through the "cLHS" R package (Roudier et al. 2012). According to Minasny and McBratney (2006), the cLHS approach is more effective to reproduce the distribution of variables than random sampling and equal spatial strata. From here onwards, the 20% validation data is referred to as "validation 01" (Fig. 1e) because we collected an additional validation data of 1,278 observations, which were collected at a farm (i.e., farm 05) outside the area where the 58,481 observations were retrieved, but inside the study site (Fig. 1d and f). From here on, this second validation dataset is referred to as "validation 02" (Fig. 1f). Validation 02 was excluded from the calibration data to prevent overoptimistic conclusions (i.e., Farm 05). All dataset were collected in the same year to remove any doubts on this proposed modelling framework.

For the calibration of SOC$_{\text{stocks}}$ models using RF, the data from farms 01 to 06 were used by applying the ten-fold cross-validation method, executed ten times, selecting the model that provided the lowest root mean square error (RMSE) and the highest model efficiency coefficient (MEC) (Janssen and Heuberger 1995). In order to validate and avoid over-optimistic conclusions, SOC$_{\text{stocks}}$ data from farm 07 were not included in the calibration and were used as external validation due to the limited number of samples.

The RMSE, MEC, concordance Lin's concordance correlation coefficient (CCC) and Bias were the selected metrics of model assessment using the validation data, as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$MEC = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu_{obs})^2} \tag{7}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{8}$$

$$CCC = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma^2_{pred} + \sigma^2_{obs} + \left(\mu_{pred} - \mu_{obs}\right)^2} \tag{9}$$

Where: n, y_i, $\hat{y}_i$, σ^2_{pred}, σ^2_{obs}, μ_{pred}, μ_{obs}, and ρ are, respectively, the sample sizes, observed values, predicted values of the response variable, the prediction and observation variances, the means of the predicted and observed values, and the correlation coefficient between the predicted and observed values.

Results

ECa and SOC$_{\text{stocks}}$ datasets

The descriptive statistics were calculated for ECa and SOC$_{\text{stocks}}$ (Table 3). ECa presented skewness and kurtosis values close to zero, as well as a high coefficient of variation. The skewness and kurtosis values for SOC$_{\text{stocks}}$ were also close to 0. The level of dispersion around the mean was higher. The calibration and validation data from 58,481 observations were split using the conditioned Latin Hypercube Sampling

Table 3 Summary statistics

Parameters	n	Min.	1st Q.	Median	Mean	3rd Q.	Max.	Skewness	Kurtosis	SD	CV
ECa_{cal}	46,784	1.40	39.40	64.50	69.29	95.00	237.60	0.53	-0.47	34.52	49.82
ECa_{cv}	11,697	6.40	39.50	65.20	69.86	95.40	235.20	0.53	-0.49	34.86	49.89
ECa_{hv}	1,278	28.50	84.30	108.20	106.10	127.50	195.40	-0.11	-0.54	29.40	27.70
SOC_{cal}	40	3.44	11.35	16.74	18.68	23.59	45.64	0.70	-0.18	10.17	54.45
SOC_{hv}	9	8.86	10.78	13.24	16.55	19.30	28.46	0.57	-1.46	7.58	45.80

*ECa_{cal}, ECa_{cv}, ECa_{hv}, SOC_{cal}, SOC_{hv} – calibration, validation 01(cross-validation), and validation 02 (hold-out validation; Farm 05) data for the apparent electrical conductivity; and calibration, and validation (hold-out validation; Farm 07) for soil organic carbon stocks

for unbiased selection of points to calibrate the models as well as points located in the same areas. This means that there was a set of validation points close to the selected calibration points (Fig. 1e). Nevertheless, we did not include one farm in the calibration data (Fig. 1f) to avoid overoptimistic conclusions in both datasets (ECa, farm 05; SOC_{stocks}, farm 07), representing a field survey after modelling ECa and SOC_{stocks}.

Modelling ECa

The Pearson's correlation analysis was computed to establish the degree of correlations between ECa and environmental variables to further describe their interrelationships (Fig. 2). ECa presented moderate correlations with the DEM, NDVI, and RENDVI from 16th of July, NDVI, RENDVI, and brightness indices from 21st of August. On the other hand, the remaining environmental variables showed weak correlation with the ECa ranging between −0.32 and 0.35.

After generating the models applying the random forest algorithm, we assessed the model performance using validation data 01 and 02 (Fig. 3). Regarding validation 01, Scenario 1 (final map named as ECa_1) fitted the ECa model showing RMSE, MEC, CCC, and Bias of 7.15 mS m^{-1}, 0.95, 0.98, and 0.05, respectively. In Scenario 2 (final map named as ECa_2), we combined environmental variables from Scenario 1 with terrain derivatives to predict ECa. This procedure improved the predictive power of the models (RMSE = 6.41 mS m^{-1}, MEC = 0.96, CCC = 0.98, and Bias = 0.09), effectively observed through the RMSE. Therefore, Scenario 2 decreased the RMSE value by 10% compared to Scenario 1. Scenario 3 (final map named as ECa_3) showed better efficacy than the other two scenarios (RMSE = 5.42 mS m^{-1}, MEC = 0.98, CCC = 0.99, and Bias = -0.01) using spectral indices calculated from RapidEye satellite collection combined with terrain derivatives from LiDAR sensor and universal kriging (Fig. S1).

Regarding validation 02 (Figs. 1d and 3) and comparing again all scenarios at once, Scenario 3 outperformed Scenarios 1 and 2. We avoided interference of natural factors, such as groundwater level, cover vegetation, and temperature changes, using calibration and validation data from the same period. The final predicted maps were generated for Scenarios 1 (Fig. 4a), 2 (Fig. 4b), and 3 (Fig. 4c).

Evaluating ECa maps and their relationship with SOC_{stocks}

The level of correlation through an exploratory analysis between SOC_{stocks} and ECa are shown in Fig. 5. Predicted ECa_2 and ECa_3 were the only environmental variables that showed a high correlation value (r = 0.78). DEM, the brightness index from 3rd of June, and the brightness index from 4th of October presented respectively r values of -0.67, -0.54, and −0.44, which mean a modest or moderate correlation. The other environmental variables displayed r values ranging from −0.36 to 0.33, categorised as having a low or weak correlation with SOC_{stocks}.

Case study: SOC_{stocks} modelling in peatlands using ECa

For case study, we modelled SOC_{stocks} comparing the level of ECa contribution as a new environmental variable, which was achieved by modelling the response variable (e.g., SOC_{stocks}) using only environmental variables from Scenario 2 and these variables

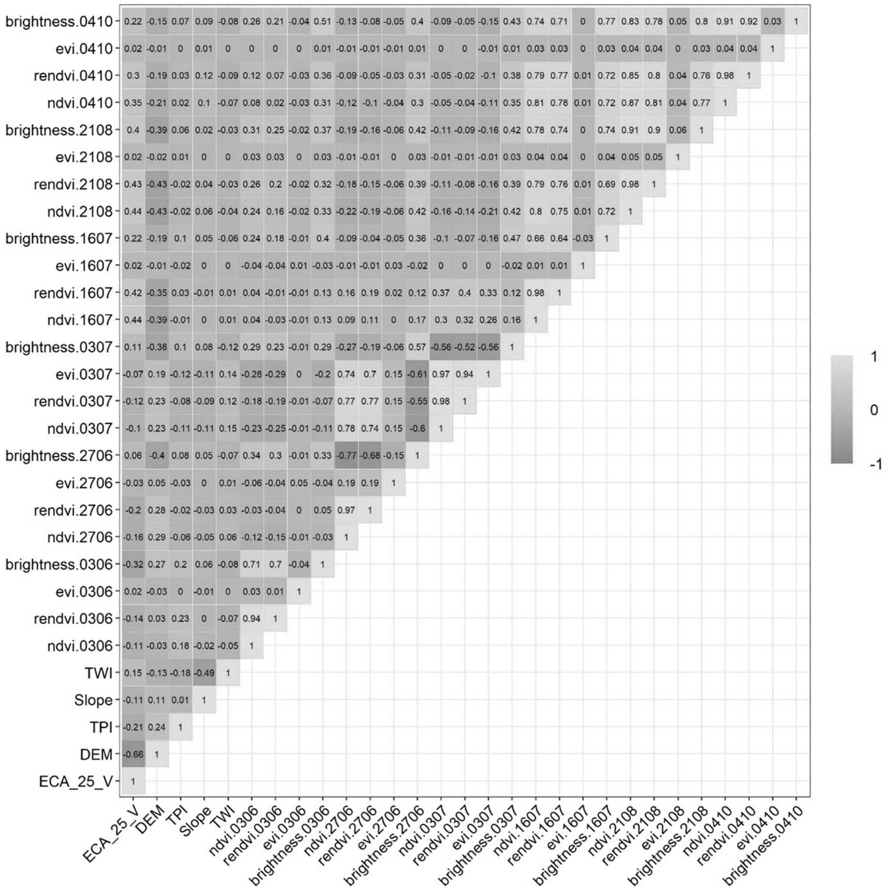

Fig. 2 Pearson's correlation coefficient between the apparent electrical conductivity and the environmental variables in the study area. The variables are represented by abbreviations: ECA_25_V, apparent electrical conductivity from electromagnetic induction instrument; DEM, digital elevation model; TPI, topographic position index; TWI, topographic wetness index; NDVI, normalised difference vegetation index; RENDVI, red-edge normalised difference vegetation index; EVI, enhanced vegetation index. The numbers correspond to the date/month format. For example, NDVI.1607 represents the normalised difference vegetation index for the 16th of July

including the best ECa predicted maps from Scenario 2 (ECa_2) and 3 (ECa_3) (Table 4 and Fig. S2).

Here, this exercise was aimed at showing the potential use of ECa maps generated by the machine learning procedure to better extrapolate SOC_{stocks} data. To ensure a robust validation of RF model with a limited dataset, we left one area out to avoid overoptimistic results. In this sense, we evaluated the potential of adding ECa information to improve the modelling predictive power of SOC_{stocks} (Table 4). Overall, the best fitted models were obtained by adding ECa information (RMSE: 4.54 and 4.51 kg m^{-2}) against the model using only environmental variables from Scenario 2 (RMSE: 5.21 kg m^{-2}). MEC also increased by ~83 and 70% for the models combining ECa with environmental variables from Scenario 2. This analysis showed that ECa can improve the modelling predictive power of SOC_{stocks}. Notably, all modelling scenarios exhibited relatively similar variable importance (Fig. 6); however, the addition of predicted ECa significantly changed the magnitude of importance.

The final maps predicted for SOC_{stocks} are shown in Fig. 7. The need to incorporate geophysics (ECa), as a tool to analyse soils, becomes evident when comparing the approaches in their spatially pattern. These spatial domains show the largest peat thickness thus the most vulnerable landscape elements in terms of drainage and related CO_2 source strength. The model strongly overestimates SOC_{stocks} without ECa information in the central part of our study site where the spatial domains are located.

Discussion

The skewness, kurtosis, and SD of the ECa dataset can be considered of normal distribution. For instance, SD in our study was 15% lower than the results found in Koszinski et al. (2015) for the same

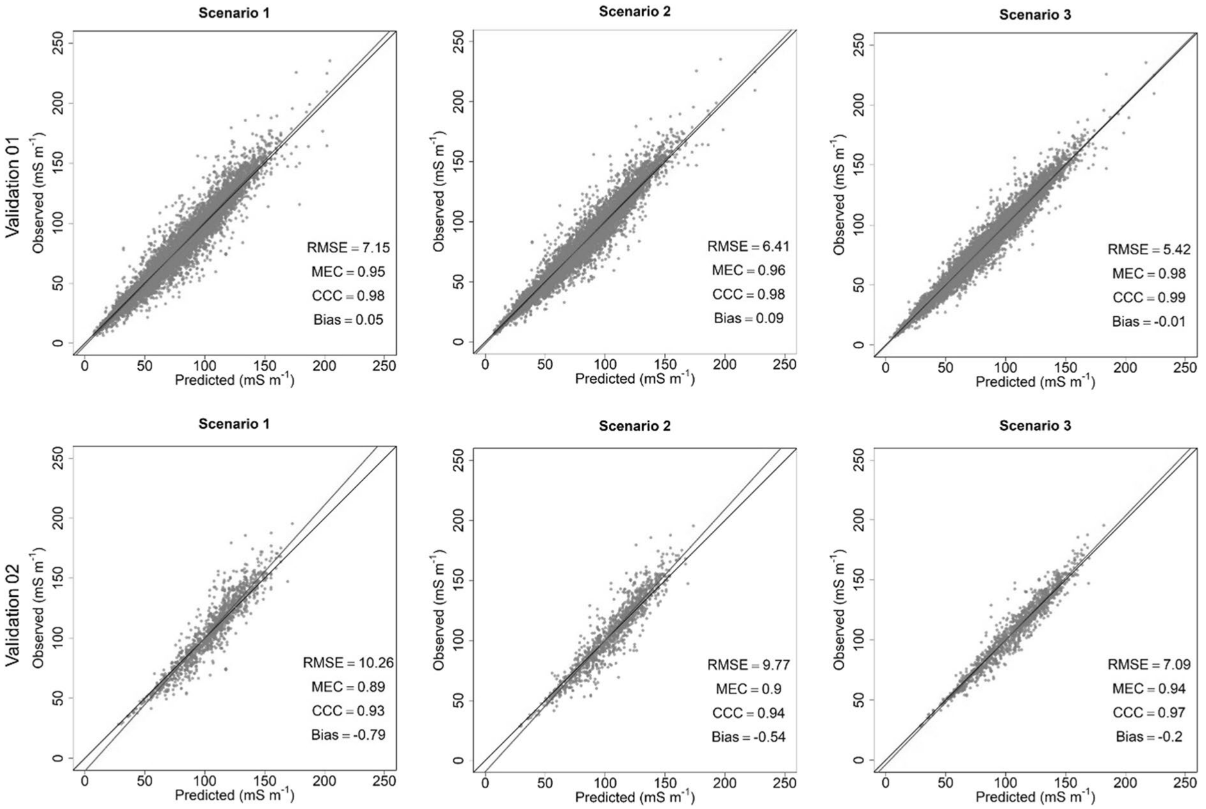

Fig. 3 Graphs of observed and predicted values for the soil apparent electrical conductivity based on both external validation dataset in the study area. Scenario 1, only spectral indices calculated from RapidEye satellite collection and random forest; Scenario 2, spectral indices calculated from RapidEye satellite collection combined with terrain derivatives from LiDAR sensor and random forest; and Scenario 3 spectral indices calculated from RapidEye satellite collection combined with terrain derivatives from LiDAR sensor and universal kriging

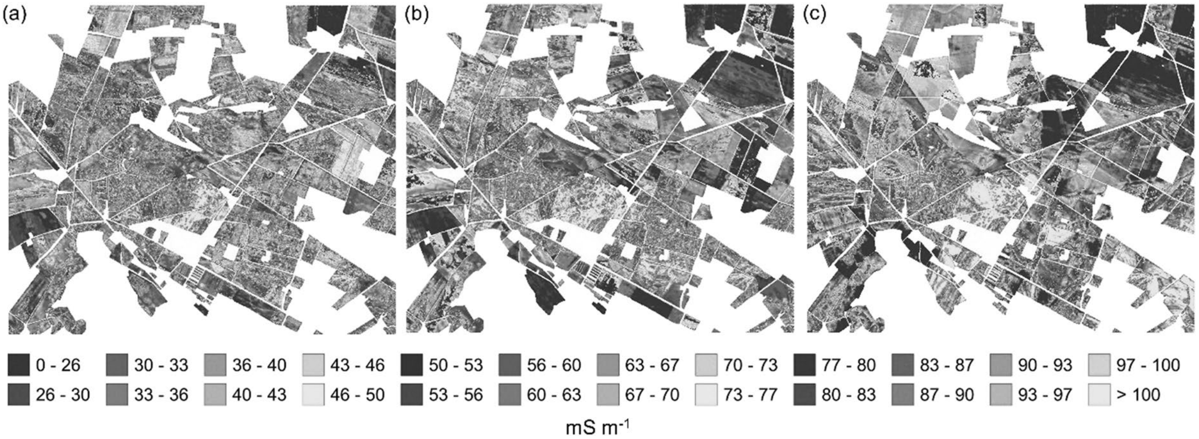

Fig. 4 Predicted maps for soil apparent electrical conductivity (ECa) from electromagnetic induction (EMI) instrument applying scenarios 1 (**a**), 2 (**b**), and 3 (**c**) in the study area

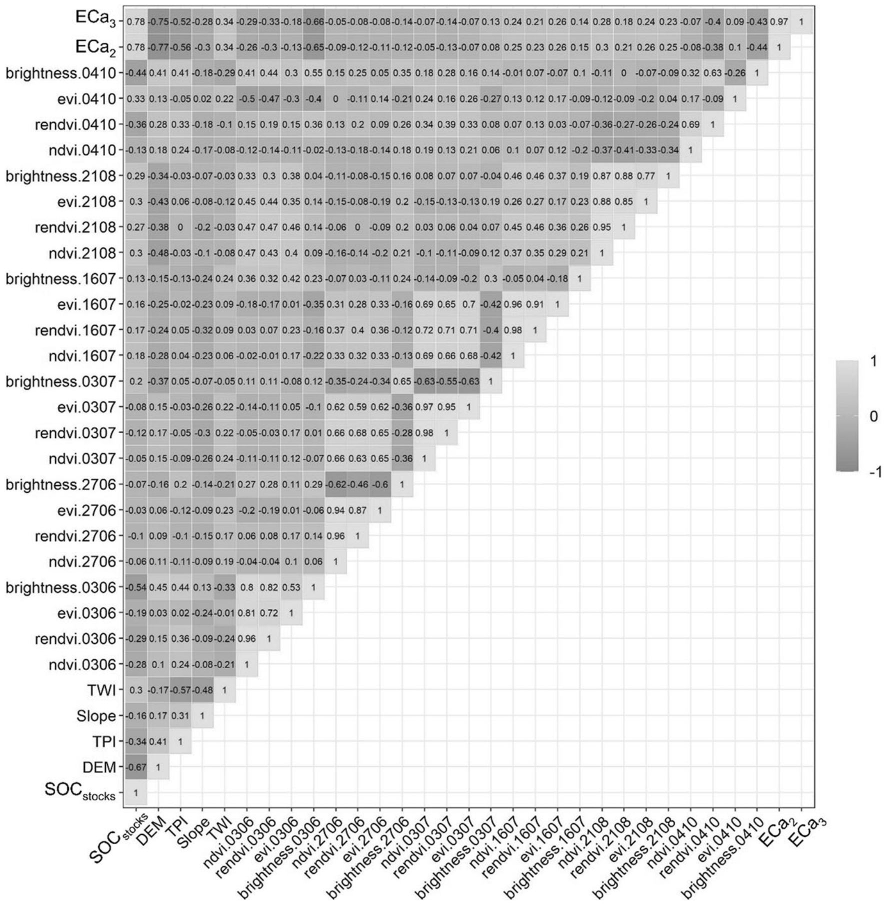

Fig. 5 Pearson correlation coefficient between the soil organic carbon stocks (SOC$_{stocks}$) and the environmental variables including the predicted ECa maps in the study area. The variables are represented by abbreviations: SOC$_{stocks}$, soil organic carbon stocks; DEM, digital elevation model; TPI, topographic position index; TWI, topographic wetness index; NDVI, normalised difference vegetation index; RENDVI, red-edge normalised difference vegetation index; EVI, enhanced vegetation index; ECa$_2$, predicted ECa map from scenario 2; ECa$_3$, predicted ECa map from scenario 3. The numbers correspond to the date/month format. For example, NDVI.1607 represents the normalised difference vegetation index for the 16th of July

Table 4 Model assessment of the predicted soil organic carbon stocks (kg m^{-2}) through Random Forest using hold-out validation data from farm 07

Models	RMSE	MEC	CCC	Bias
Environmental variables from Scenario 2	5.21	0.43	0.63	-0.43
ECa$_2$ with the environmental variables from Scenario 2	4.54	0.79	0.75	2.60
ECa$_3$ with the environmental variables from Scenario 2	4.51	0.73	0.75	1.50

study site. SOC$_{stocks}$ presented a normal distribution based on the descriptive statistics. The SD value for SOC$_{stocks}$ was also lower (~28%) than that described in Koszinski et al. (2015). The normal distribution is a desirable characteristic in the response variable for modelling, which is unlikely to occur in a natural environment (Malone et al. 2013). However, we found a normal distribution for ECa and SOC$_{stocks}$.

Two validation data were performed for ECa in this study (Fig. 1c and d). The first was selected by splitting the 58,481 observations into 80/20% for calibration and validation, respectively. Validation 01 had observations in the same sites of the calibration set (Fig. 1e). The procedure applied in that case was the conditioned Latin Hypercube Sampling (cLHS), which split observations based on environmental variables for a more effective model training of the response variable (Brungard and Boettinger 2010). Furthermore, cLHS was recommended by Minasny et al. (2019) reviewing the current state of digital mapping of peatlands, requiring further research. Validation 02 (i.e., farm 05) was composed of 1,278

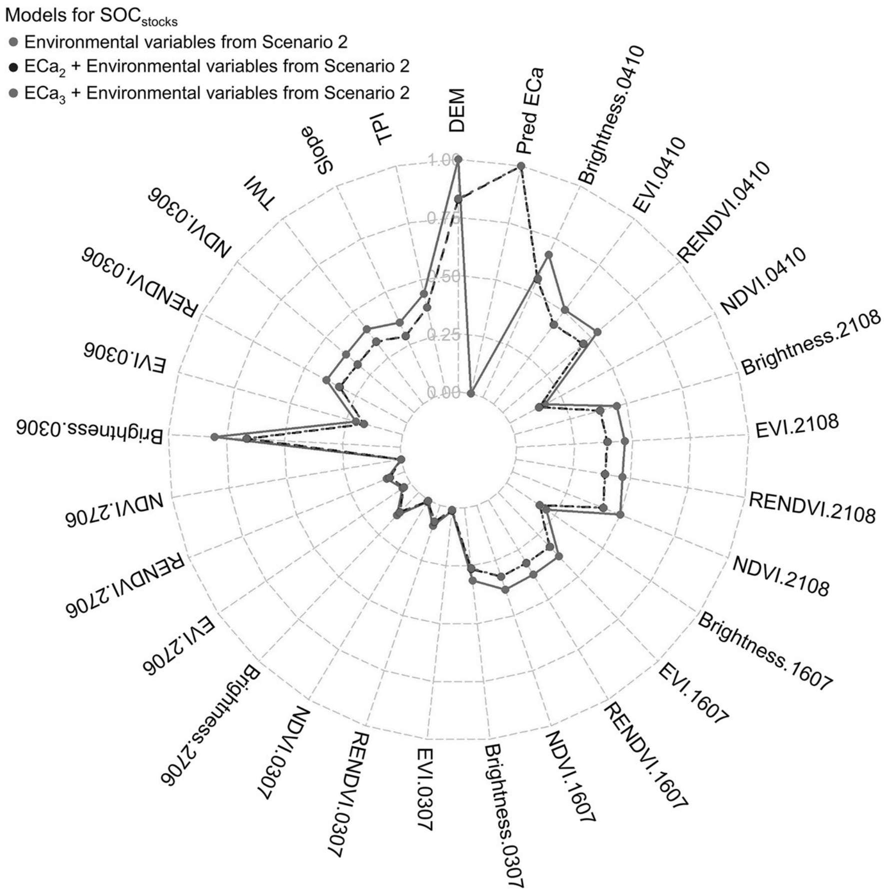

Fig. 6 Relative contribution values of environmental variables in modelling SOC$_{stocks}$ using Random Forest. The variables are represented by abbreviations: DEM, digital elevation model; TPI, topographic position index; TWI, topographic wetness index; NDVI, normalised difference vegetation index; RENDVI, red-edge normalised difference vegetation index; EVI, enhanced vegetation index; Pred ECa, predicted ECa map from scenarios 2 (ECa$_2$) and 3 (ECa$_3$). The numbers correspond to the date/month format. For example, NDVI.1607 represents the normalised difference vegetation index for the 16th of July

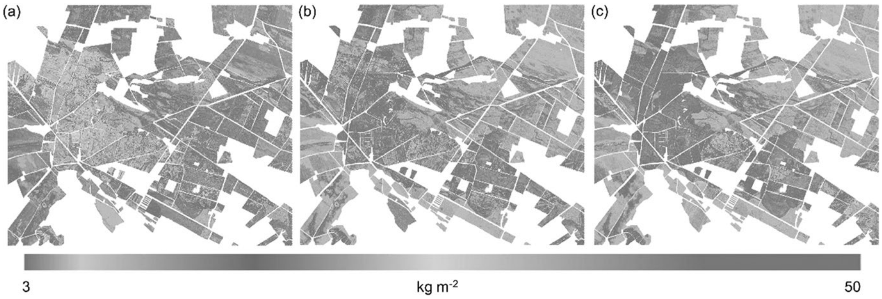

Fig. 7 Predicted maps of SOC$_{stocks}$ using only environmental variables from scenario 2 (**a**), with predicted ECa applying random forest, and (**c**) predicted ECa applying Universal Kriging

observations (Fig. 1f), which could represent a new field survey after generating ECa predicted maps. Moreover, we avoided unbiased evaluation of the models using the second validation data. This procedure was crucial to represent the real *in-situ* situation to consistently determine the well-fitted models among the modelling scenarios.

Six environmental variables had the highest correlation coefficient with ECa, although that correlation was considered moderate (Mason et al. 1983). Notably, among these variables, DEM showed the highest association with ECa (Fig. 2). In a similar context, Koszinski et al. (2015) reported a Pearson's correlation coefficient of -0.69 between ECa_v and DEM. The expected negative correlations between DEM and ECa underscore the need to address these influences when interpreting ECa mapping results (Figs. 3 and 4). The topographical features captured in the DEM can impact water flow, drainage patterns, and soil moisture levels, which, in turn, affect the electrical conductivity of the subsurface. Therefore, low-lying areas (e.g., depressions and valleys) may accumulate more moisture, leading to higher electrical conductivity values as water enhances the conductivity of the soil. Conversely, higher elevations (e.g., hills and ridges) might have lower moisture content, resulting in lower electrical conductivity values.

The assessment of model performance was conducted using both validation data after generating the models. Overall, the best fitted models for ECa were Scenarios 2 and 3, respectively, performing RF (RMSE: 6.41 and 9.77 mS m^{-1}; MEC: 0.96 and 0.90) and universal kriging (RMSE: 5.42 and 7.09 mS m^{-1}; MEC: 0.98 and 0.94). Taghizadeh-Mehrjardi et al. (2014) found RMSE of 37.74 mS m^{-1} and MEC of 0.49 mapping ECa using the local regression kriging approach and terrain derivatives from Landsat ETM + as environmental variables. Wu et al. (2018) predicted soil salinity (ECa) using Landsat ETM + and ALOS data as environmental variables through RF and Support Vector Machine algorithms. The authors found MEC values between 0.72 and 0.89. Therefore, our results presented better accuracy than most studies conducted to date (Saey et al. 2012; Yang et al. 2019; Zhang et al. 2020). We attribute the high accuracy of our models to high-resolution digital elevation model and its derivatives, as well as the high-resolution multispectral data from RapidEye satellite sensor in the same year of ECa field collection. The final predicted maps of ECa showed high values from the NW-SE direction, which corroborated the geological formation related to glacial processes during the Pleistocene phase (Koszinski et al. 2015). The microrelief features in the region are also clear. The study site is characterised by extensive floodpains and bas-relief; however, there are small spots with high ECa concentration from summit to toeslope in the NE and SW directions. This shows the importance of predicted ECa maps to characterise peat thickness and help to extrapolate information to unknown areas, which can be achieved by applying machine learning and geostatistical methods, such as RF and UK.

The importance and novelty of our procedure to generate ECa maps lie in the fact that ECa is highly correlated with SOC_{stocks}. Altdorff et al. (2016) found the Pearson's correlation (r) values ranging from 0.4 to 0.3 between ECa and SOC_{stocks} at three different soil peat depths (25, 50, and 100 cm). Koszinski et al. (2015) also achieved r value of 0.79 for information between ECa and SOC_{stocks}. Thus, these studies corroborate our findings and highlight the high potential of ECa to quantify and extrapolate SOC_{stocks} information in peatlands. The high correlation among the mapped ECa_2, ECa_3 and SOC_{stocks} prove the efficacy of ECa modelling through RF and UK approaches (Fig. 5). Our study demonstrates that integrating spatially explicit ECa maps as engineered variables significantly enhances the accuracy of predicting SOC_{stocks}. This departure from solely relying on base variables highlights the value of exploring new environmental variables to improve model calibration and prediction. Additionally, we observed significant improvements in SOC_{stocks} predictions with the inclusion of ECa data (Fig. S2 and Table 4), further emphasising the pivotal role of geophysical information (Fig. 6). The engineering of new variables, represented by spatially explicit ECa maps, proves to be highly beneficial in predicting the SOC_{stocks}.

Questions about the advantages of using machine learning or universal kriging models may arise. Therefore, performing UK and RF on ECa data allowed mapping the total area of 26.54 km^2 rather than only 1.12 km^2. ECa data could be extrapolated 23.7 times using RF and UK modelling. This study directs further research into extrapolating and replicating this methodology through high-resolution digital elevation models and their derivatives, as well as the high-resolution multispectral data from RapidEye satellite sensor in the same year of ECa field collection. Other sources of high-resolution multispectral data should be evaluated taking into account different continental climates and ecosystems. This shows the potential of using machine learning approaches to predict ECa from EMI sensors whether spatial, temporal, spectral, and sampling components in the study site.

Another question raised regards how large a dataset should be to fit a machine learning, which contained 49 observations for SOC_{stocks} in our study. However, our validation strategy leaving one area out (farm 07: 9 observations) could provide valuable information on the efficacy of RF models and avoid overoptimistic conclusions. According to Padarian et al. (2020), there is no clear rule for the dataset size that constitutes a problem to use machine learning methods. It depends on how complex the problem is and its relationship with the environmental variables that can help explain or understand the problem. Therefore, ECa and high-resolution remote sensing data improved the prediction power of SOC_{stocks} modelling in peatlands.

The drainage network and groundwater patterns can be clearly detected by the modelling scenarios (Fig. 4). It is vital to detect these characteristics in peatlands because agricultural lands involve draining peat, while increasing peat consolidation and decomposition (Hoogland et al. 2012), leading to land subsidence of peat soils (Minasny et al. 2019). Thus, in case there is a slight chance of extrapolating ECa information to better detect the aforementioned characteristics, it has to be taken into account. Therefore, we predicted SOC_{stocks} with and without ECa information as one of the environmental variables (Fig. 7). Our results show that ECa improved SOC_{stocks} predictions compared with predictions without ECa, evidencing the need to incorporate geophysics (ECa) into methods as a tool to analyse soils. The resulting map reveals important insights into the distribution of SOC_{stocks} across the study area. We observe distinct spatial patterns, including areas with high and low SOC_{stocks}, which can be attributed to different environmental factors, land management practices, and hydrological conditions. Proximal sensing data, such as that predicted ECa, is one of the most important environmental variables to enhance the mapping of peatlands and their properties (i.e., SOC_{stocks}).

Conclusions

The methodology and approach used in this study showed the feasibility of generating ECa maps from field survey EMI data and utilising them as a valuable environmental variable for predicting SOC_{stocks} in peatlands. Through extensive model assessments, our models exhibited higher accuracy compared to previous studies, attributed to the integration of spatially explicit ECa maps as engineered variables that significantly enhanced the accuracy of predicting SOC_{stocks}, emphasising the importance of exploring new environmental variables in digital soil mapping.

Our study also highlights the scalability and benefits of machine learning approaches for predicting ECa from EMI sensors across larger areas. While the dataset size for machine learning methods remains a subject of consideration, our validation strategy effectively assessed the model performance and avoided overoptimistic conclusions. Overall, our findings indicate that ECa, along with high-resolution remote sensing data, can enhance the prediction power of SOC_{stocks} modelling in peatlands. We advocate for the incorporation of proximal sensing data, such as predicted ECa from field survey EMI data, as a critical environmental variable in optimising peatland mapping and property characterisation, surpassing the reliance solely on LiDAR data.

We hope our research opens new avenues for further investigations, harnessing the potential of machine learning approaches, proximal and remote sensing to engineer new environmental variables. This advancement enhances our understanding and management of peatland soils and other terrestrial ecosystems.

Acknowledgements The authors would like to thank Norbert Wypler and Wilfried Hierold for extensive field work and expert knowledge in soil sampling and description. We deeply recognise the efforts made by the anonymous referees to markedly improve the quality of the article.

Author contribution All authors contributed to the study conception, design, material preparation, data collection and analysis. The first draft of the manuscript was written by Wanderson de Sousa Mendes and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding for this study was provided by the German Federal Ministry of Education and Research (BMBF, project: “Climate protection by peatland protection – Strategies for peatland management”, 01LS05049).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Abdu H, Robinson DA, Jones SB (2007) Comparing bulk soil electrical conductivity determination using the DUALEM-1S and EM38-DD electromagnetic induction instruments. Soil Sci Soc Am J 71:189–196. https://doi.org/10.2136/sssaj2005.0394

Agbu PA, Fehrenbacher DJ, Jansen IJ (1990) Soil property relationships with SPOT satellite digital data in East Central Illinois. Soil Sci Soc Am J 54:807. https://doi.org/10.2136/sssaj1990.03615995005400030031x

Altdorff D, Bechtold M, van der Kruk J, Altdorff D, Bechtold M, van der Kruk J, Vereecken H, Huisman JA (2016) Mapping peat layer properties with multi-coil offset electromagnetic induction and laser scanning elevation data. Geoderma 261:178–189. https://doi.org/10.1016/j.geoderma.2015.07.015

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Brevik EC, Fenton TE, Horton R (2004) Effect of daily soil temperature fluctuations on soil slectrical conductivity as measured with the Geonics® EM-38. Precis Agric 2004 52 5:145–152. https://doi.org/10.1023/B:PRAG.0000022359.79184.92

Brungard CW, Boettinger JL (2010) Conditioned latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. Digital Soil Mapping. Springer Netherlands, Dordrecht, pp 67–75

Christensen R (2011) Plane answers to complex questions, 4th edn. Springer New York, New York, NY

Comas X, Slater L (2004) Low-frequency electrical properties of peat. Water Resour Res 40:1–9. https://doi.org/10.1029/2004WR003534

DIN EN ISO 11272 2014-06 (2014) Soil quality - Determination of dry bulk density (ISO 11272:1998). Beuth Verlag GmbH: Berlin, Germany, 1–15. https://www.beuth.de/en/standard/din-en-iso-11272/203694686

DIN ISO10694 1996-08 (1996) Soil quality - Determination of organic and total carbon after dry combustion (elementary analysis) (ISO 10694:1995). Beuth Verlag GmbH: Berlin, Germany, 1–5. https://www.beuth.de/en/standard/din-iso-10694/2799936

Forkuor G, Hounkpatin OKL, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple Linear regression models. PLoS ONE 12:e0170478. https://doi.org/10.1371/journal.pone.0170478

Gianola D, Okut H, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet 12:87. https://doi.org/10.1186/1471-2156-12-87

Gitelson A, Merzlyak MN (1994) Spectral reflectance changes associated with autumn senescence of aesculus hippocastanum L. and acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation. J Plant Physiol 143:286–292. https://doi.org/10.1016/S0176-1617(11)81633-0

Gorham E (1991) Northern Peatlands: role in the carbon cycle and probable responses to climatic warming. Ecol Appl 1:182–195. https://doi.org/10.2307/1941811

Grinand C, Le Maire G, Vieilledent G et al (2017) Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing. Int J Appl Earth Obs Geoinf 54:1–14. https://doi.org/10.1016/j.jag.2016.09.002

Hartemink AE, Zhang Y, Bockheim JG, et al (2020) Soil horizon variation: A review. Adv Agron 160:125–185. https://doi.org/10.1016/bs.agron.2019.10.003

Hoogland T, van den Akker JJH, Brus DJ (2012) Modeling the subsidence of peat soils in the dutch coastal area. Geoderma 171–172:92–97. https://doi.org/10.1016/j.geoderma.2011.02.013

Houborg R, McCabe MF (2018) A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. ISPRS J Photogramm Remote Sens 135:173–188. https://doi.org/10.1016/j.isprsjprs.2017.10.004

Huang J, Monteiro Santos FA, Triantafilis J (2016) Mapping soil water dynamics and a moving wetting front by spatiotemporal inversion of electromagnetic induction data. Water Resour Res 52:9131–9145. https://doi.org/10.1002/2016WR019330

Huete AR (2004) Remote sensing for environmental monitoring. Environ Monit Charact 183–206. https://doi.org/10.1016/B978-012064477-3/50013-8

IUSS Working Group WRB (2015) World reference base for soil resources 2014: international soil classification system for naming soils and creating legends for soil maps. FAO, Rome

Jackson RB, Lajtha K, Crow SE, Jackson RB, Lajtha K, Crow SE, Hugelius G, Kramer MG, Piñeiro G (2017) The Ecology of soil carbon: pools, vulnerabilities, and biotic and abiotic controls. Annu Rev Ecol Evol Syst 48:419–445. https://doi.org/10.1146/annurev-ecolsys-112414-054234

Janssen PHM, Heuberger PSC (1995) Calibration of process-oriented models. Ecol Modell 83:55–66. https://doi.org/10.1016/0304-3800(95)00084-9

Keskin H, Grunwald S (2018) Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326:22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

Kettridge N, Comas X, Baird A et al. (2008) Ecohydrologically important subsurface structures in peatlands revealed by ground-penetrating radar and complex conductivity surveys. J Geophys Res Biogeosci 113. https://doi.org/10.1029/2008JG000787

Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. Appl Math Model 81:401–418. https://doi.org/10.1016/j.apm.2019.12.016

Koszinski S, Miller BA, Hierold W, Koszinski S, Miller BA, Hierold W, Haelbich H, Sommer M (2015) Spatial modeling of organic carbon in degraded Peatland soils of Northeast Germany. Soil Sci Soc Am J 79:1496–1508. https://doi.org/10.2136/sssaj2015.01.0019

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26. https://doi.org/10.18637/jss.v028.i05

Lamichhane S, Kumar L, Wilson B (2019) Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. Geoderma 352:395–413

Lees KJ, Quaife T, Artz RRE, Lees KJ, Quaife T, Artz RRE, Khomik M, Clark JM (2018) Potential for using remote sensing to estimate carbon fluxes across northern peatlands – A review. Sci Total Environ 615:857–874

Loisel J, van Bellen S, Pelletier L, Loisel J, van Bellen S, Pelletier L, Talbot J, Hugelius G, Karran D, Yu Z, Nichols J, Holmquist J (2017) Insights and issues with estimating northern peatland carbon stocks and fluxes since the last glacial Maximum. Earth Sci Rev 165:59–80. https://doi.org/10.1016/j.earscirev.2016.12.001

Malone BP, McBratney AB, Minasny B (2013) Spatial scaling for digital soil mapping. Soil Sci Soc Am J 77:890–902. https://doi.org/10.2136/sssaj2012.0419

Mason RO, Lind DA, Marchal WG (1983) Statistics: an introduction. Harcourt Brace Jovanovich, Inc, New York

McBratney AB, Mendonça Santos ML, Minasny B (2003) On digital soil mapping. Geoderma 117:3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Miller BA, Koszinski S, Hierold W, Miller BA, Koszinski S, Hierold W, Rogasik H, Schröder B, Van Oost K, Wehrhan M, Sommer M (2016) Towards mapping soil carbon landscapes: issues of sampling scale and transferability. Soil Tillage Res 156:194–208. https://doi.org/10.1016/j.still.2015.07.004

Minasny B, Berglund Ö, Connolly J, Minasny B, Berglund Ö, Connolly J, Hedley C, de Vries F, Gimona A, Kempen B, Kidd D, Lilja H, Malone B, McBratney A, Roudier P, O'Rourke S et al (2019) Digital mapping of peatlands – a critical review. Earth Sci Rev 196:102870. https://doi.org/10.1016/j.earscirev.2019.05.014

Minasny B, McBratney AB (2006) A conditioned latin hypercube method for sampling in the presence of ancillary information. Comput Geosci 32:1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009

Mueller L, Behrendt A, Shepherd TG, Mueller L, Behrendt A, Shepherd TG, Schindler U, Kaiser T (2007) Implications of soil substrate and land use for properties of fen soils in North-East Germany Part III: soil quality for grassland use. Arch Agron Soil Sci 53:137–146. https://doi.org/10.1080/03650340701224831

Padarian J, Minasny B, McBratney AB (2020) Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6:35–52. https://doi.org/10.5194/soil-6-35-2020

QGIS Development Team (2020) QGIS geographic information system. QGIS Association. http://www.qgis.org/

Roudier P, Hewitt A, Beaudette D (2012) A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Digit Soil Assessments Beyond 227–231. https://doi.org/10.1201/b12728-46

Rouse JW, Hass RH, Schell JA, Deering DW (1973) Monitoring vegetation systems in the great plains with ERTS. Third Earth Resour Technol Satell Symp 1:309–317

Saey T, Islam MM, De Smedt P, Saey T, Islam MM, De Smedt P, Meerschman E, Van De Vijver E, Lehouck A, Van Meirvenne M (2012) Using a multi-receiver survey of apparent electrical conductivity to reconstruct a Holocene tidal channel in a polder area. CATENA 95:104–111. https://doi.org/10.1016/j.catena.2012.02.001

Schlichting E, Blume H-P, Stahr K (1995) Practicing soil science. An introduction to pedological working for ecologists, especially agriculturists and for earth scientists, 2nd edn. Blackwell Publishing Ltd, Berlin

Sims DA, Gamon JA (2002) Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sens Environ 81:337–354. https://doi.org/10.1016/S0034-4257(02)00010-X

Sun W, Whelan BM, Minasny B, Mcbratney AB (2012) Evaluation of a local regression kriging approach for mapping apparent electrical conductivity of soil (ECa) at high resolution. J Plant Nutr Soil Sci 175:212–220. https://doi.org/10.1002/jpln.201100005

Taghizadeh-Mehrjardi R, Minasny B, Sarmadian F, Malone BP (2014) Digital mapping of soil salinity in ardakan region, central iran. Geoderma 213:15–28. https://doi.org/10.1016/j.geoderma.2013.07.020

Taylor R (1990) Interpretation of the correlation coefficient: a basic review. J Diagn Med Sonogr 6:35–39. https://doi.org/10.1177/875647939000600106

Thompson JA, Kienast-Brown S, D'Avello T et al (2020) Soils2026 and digital soil mapping – a foundation for the future of soils information in the United States. Geoderma Reg 22:e00294. https://doi.org/10.1016/j.geodrs.2020.e00294

Tiemeyer B, Albiac Borraz E, Augustin J, Tiemeyer B, Albiac Borraz E, Augustin J, Bechtold M, Beetz S, Beyer C, Drösler M, Ebli M, Eickenscheidt T, Fiedler S, Förster C, Freibauer A, Giebels M, Glatzel S, Heinichen J, Hoffmann M, Höper H, Jurasinski G, Leiber-Sauheitl K, Peichl-Brak M, Roßkopf N, Sommer M, Zeitz J (2016) High emissions of greenhouse gases from grasslands on peat and other organic soils. Glob Chang Biol 22:4134–4149. https://doi.org/10.1111/gcb.13303

Tiemeyer B, Freibauer A, Borraz EA, Tiemeyer B, Freibauer A, Borraz EA, Augustin J, Bechtold M, Beetz S, Beyer C, Ebli M, Eickenscheidt T, Fiedler S, Förster C, Gensior A, Giebels M, Glatzel S, Heinichen J, Hoffmann M, Höper H, Jurasinski G, Laggner A, Leiber-Sauheitl K, Peichl-Brak M, Drösler M (2020) A new methodology for organic soils in national greenhouse gas inventories: data synthesis, derivation and application. Ecol Indic 109:105838. https://doi.org/10.1016/j.ecolind.2019.105838

Trangmar BB, Yost RS, Uehara G (1986) Application of geostatistics to spatial studies of soil properties. Adv Agron 38:45–94. https://doi.org/10.1016/S0065-2113(08)60673-2

Wadoux AMJC, Heuvelink GBM, de Bruin S, Brus DJ (2021) Spatial cross-validation is not the right way to evaluate map accuracy. Ecol Modell 457:1–5. https://doi.org/10.1016/J.ECOLMODEL.2021.109692

Wadoux AMJCMJ-C, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci Rev 210:103359. https://doi.org/10.31223/OSF.IO/8EQ6S

Walter J, Lück E, Bauriegel A et al (2015) Multi-scale analysis of electrical conductivity of peatlands for the assessment of peat properties. Eur J Soil Sci 66:639–650. https://doi.org/10.1111/ejss.12251

Weiss A (2001) Topographic position and landforms analysis. In: ESRI Users Conference. San Diego, CA

Wiesmeier M, Urbanski L, Hobley E, Wiesmeier M, Urbanski L, Hobley E, Lang B, von Lützow M, Marin-Spiotta E, van Wesemael B, Rabot E, Ließ M, Garcia-Franco N, Wollschläger U, Vogel H-J, Kögel-Knabner I (2019) Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma 333:149–162

Wilson J, Gallant J (2000) Secondary topographic attributes. In: Wilson, JP; Gallant, JC eds, editor/s. Terrain Analysis: Principles and Applications. New York: John Wiley and Sons 87–131

Wu W, Zucca C, Muhaimeed AS et al (2018) Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq. Land Degrad Dev 29:4005–4014. https://doi.org/10.1002/ldr.3148

Yang SH, Liu F, Song XD, Yang S-H, Liu F, Song X-D, Lu Y-Y, Li D-C, Zhao Y-G, Zhang G-L (2019) Mapping topsoil electrical conductivity by a mixed geographically weighted regression kriging: a case study in the Heihe River Basin, northwest China. Ecol Indic 102:252–264. https://doi.org/10.1016/j.ecolind.2019.02.038

Yu Z, Loisel J, Brosseau DP et al (2010) Global peatland dynamics since the last glacial Maximum. Geophys Res Lett 37:1–5. https://doi.org/10.1029/2010GL043584

Zhang Y, Ji W, Saurette DD, Zhang Y, Ji W, Saurette DD, Easher TH, Li H, Shi Z, Adamchuk VI, Biswas A (2020) Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma 366:114253. https://doi.org/10.1016/j.geoderma.2020.114253

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.