remote sensing

MDPI

Article

Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils

Tom Broeg [1,2,*], Michael Blaschek [3], Steffen Seitz [2], Ruhollah Taghizadeh-Mehrjardi [2,4], Simone Zepp [5] and Thomas Scholten [2,4,6]

1 Thünen Institute of Farm Economics, Bundesallee 63, 38116 Braunschweig, Germany
2 Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, 72070 Tübingen, Germany
3 State Authority for Geology, Resources and Mining, Albertstraße 5, 79104 Freiburg, Germany
4 CRC 1070 Ressource Culture, University of Tübingen, 72070 Tübingen, Germany
5 German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Muenchener Str. 20, 82234 Wessling, Germany
6 DFG Cluster of Excellence "Machine Learning", University of Tübingen, 72070 Tübingen, Germany
* Correspondence: tom.broeg@thuenen.de

Abstract: Precise knowledge about the soil organic carbon (SOC) content in cropland soils is one requirement to design and execute effective climate and food policies. In digital soil mapping (DSM), machine learning algorithms are used to predict soil properties from covariates derived from traditional soil mapping, digital elevation models, land use, and Earth observation (EO). However, such DSM models are trained for a specific dataset and region and have so far only allowed limited general statements to be made that would enable the models to be transferred to different regions. In this study, we test the transferability of SOC models for cropland soils using five different covariate groups: multispectral soil reflectance composites (satellite), soil legacy data (soil), digital elevation model derivatives (terrain), climate parameters (climate), and combined models (combined). The transferability was analyzed using data from two federal states in southern Germany: Bavaria and Baden-Wuerttemberg. First, baseline models were trained for each state with combined models performing best in both cases (R^2 = 0.68/0.48). Next, the models were transferred and tested with soil samples from the other state whose data were not used during model calibration. Only satellite and combined models were transferable, but accuracy declined in both cases. In the final step, models were trained with samples from both states (mixed-data models) and applied to each state separately. This process significantly improved the accuracies of satellite, terrain, and combined models, while it showed no effect on climate models and decreased the models based on soil covariates. The experiment underlines the importance of EO for the transfer and extrapolation of DSM models.

Keywords: machine learning; digital soil mapping; soil organic carbon; model transfer; extrapolation; soil reflectance composite; legacy soil maps; Baden-Württemberg; Bavaria

check for updates

Citation: Broeg, T.; Blaschek, M.; Seitz, S.; Taghizadeh-Mehrjardi, R.; Zepp, S.; Scholten, T. Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils. *Remote Sens.* **2023**, *15*, 876. https://doi.org/10.3390/rs15040876

Academic Editors: Dominique Arrouays and Emmanuelle Vaudour

Received: 19 December 2022
Revised: 20 January 2023
Accepted: 31 January 2023
Published: 4 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soil organic carbon (SOC) is not only a large sink for atmospheric CO_2 but also one of the most important factors of soil quality and soil fertility [1]. Since the beginning of human land-use activities, most intensely used soils lost a large fraction of their natural SOC stock [2]. This is especially true for cropland soils which depend on good management to keep a stable carbon equilibrium [3]. In environmental policy, this deficit is commonly portrayed as an opportunity for synergetic policymaking; to use the ability of soils to store organic carbon to sequester CO_2 and enhance soil quality and food security [4]. To improve our understanding of cropland SOC stocks and support soil management, precise maps with a high spatial resolution are needed [5].

The framework of digital soil mapping (DSM) is based on the idea that soil properties are represented by environmental covariates and can be mapped with soil spatial prediction functions (SSPF) [6,7]. The basic DSM framework has been updated to account for modern machine learning (ML) algorithms [8] and was successfully used to predict SOC over large areas [9,10]. Along with terrain attributes, Earth observation (EO) data are important assets of DSM and multispectral images of bare soil are often used to predict SOC [11–13]. Most times of the year, cropland is densely covered with vegetation and the soils of different areas are rarely exposed at the same time. To solve this problem, multi-temporal EO data are analyzed to detect bare soil areas over multiple years [14–17]. Spectral indices such as the Normalized Difference Vegetation Index (NDVI) [18], the Bare Soil Index (BSI) [15], or the Normalized Burn Ratio 2 (NBR2) [18] are calculated to differentiate between covered and exposed soil and the reflectance of each pixel is averaged over a defined period. The resulting soil reflectance composite maps can be used as covariates to model SOC and other soil properties in cropland soils [16,18–20].

Ground truth is necessary to train and validate DSM models but soil samples can be scarce in some regions [21], limited in field operability [22], and analysis is expensive and time-consuming [23]. Efforts have been made to reduce the number of necessary ground measurements by extrapolating or transferring DSM models from donor areas that have a higher availability of soil data [24–31]. In most cases, models of soil types or classes were transferred with mixed overall accuracies between 46% [29] and 83% [30]. One study was able to transfer a SOC model with a relative overall accuracy of 70% [28]. The models were either based on expert knowledge [24], regressions [25], ML algorithms such as artificial neural networks (ANN) [26], or random forest (RF) [30]. The choice of method is important as some models, such as RF, are unable to extrapolate values outside the range of the training data [32]. However, a direct comparison between different model types showed that tree-based algorithms such as RF had the highest accuracies in extended modeling areas [29]. Since RF is one of the most common methods for DSM [8], this study aims to define preconditions to successfully transfer RF models to new areas.

Covariates from various sources such as remote sensing [29], ground-based detection [25], digital elevation models (DEM) [30], and soil legacy data [28] have been used to transfer machine DSM models. It is difficult to compare covariates between different studies as ML algorithms are sensitive to the context of the training dataset which often reflects a specific research area [10,33]. Furthermore, covariates reflect different data sources, and some covariates groups likely express a higher degree of transferability while others should only be used in a local context [28].

In this study, the transferability of SOC models based on multispectral soil reflectance composites (satellite), soil legacy data (soil), DEM derivatives (terrain), climate parameters (climate), as well as combined models (combined) were tested. This was done by comparing models in the two southern German federal states of Bavaria and Baden-Wuerttemberg, with a combined area of about 100,000 km^2. The degree of transferability was analyzed in three steps: (i) First, a baseline was set by testing the accuracy of each covariate group for the prediction of SOC in both states. (ii) These models were then transferred and tested with the soil samples of the respective other state to analyze how the performance changes if the models are applied to a new area. (iii) In the last step, mixed data models with different proportions of training data from both states were built to evaluate the effect of additional samples from outside the validation area. This was repeated for each covariate group and both states.

2. Materials and Methods

2.1. Study Area

The study area includes the federal states of Baden-Wuerttemberg (BW) and Bavaria (BY), in southern Germany (Figure 1A,B). Summary statistics of both states are provided in Table 1. The climate is temperate and humid with mean temperatures of about 9 °C and precipitation above 800 mm. Southern Germany is a diverse landscape and both states

share most geologic formations and parent materials of pedogenesis [34]. Most of the area of BW and the northern parts of BY is located in the Southwest German Scarplands which are characterized by fast-changing strata of the Triassic and Jurassic. BW is characterized by its distinct geological features, namely the Rhine Valley and the Black Forest in the west followed by the Swabian Alps, ranging towards the Bavarian border in the east (Figure 1A). The southeast of BW and Southern BY is covered by the North Alpine Foreland Basin which is filled with sediments of the Alps and was further formed by the glacial activities of the Pleistocene. Luvisols and cambisols are the most frequent soil types while groundwater-influenced and carbon-rich gleysols and fluvisols are connected to streams originating from the Alps [35]. Cropland is the predominant land use type throughout the research area, excluding the mountainous areas. The models in this study were limited to cropland without permanent vegetation and regularly exposed soils [14] (Figure 1C).

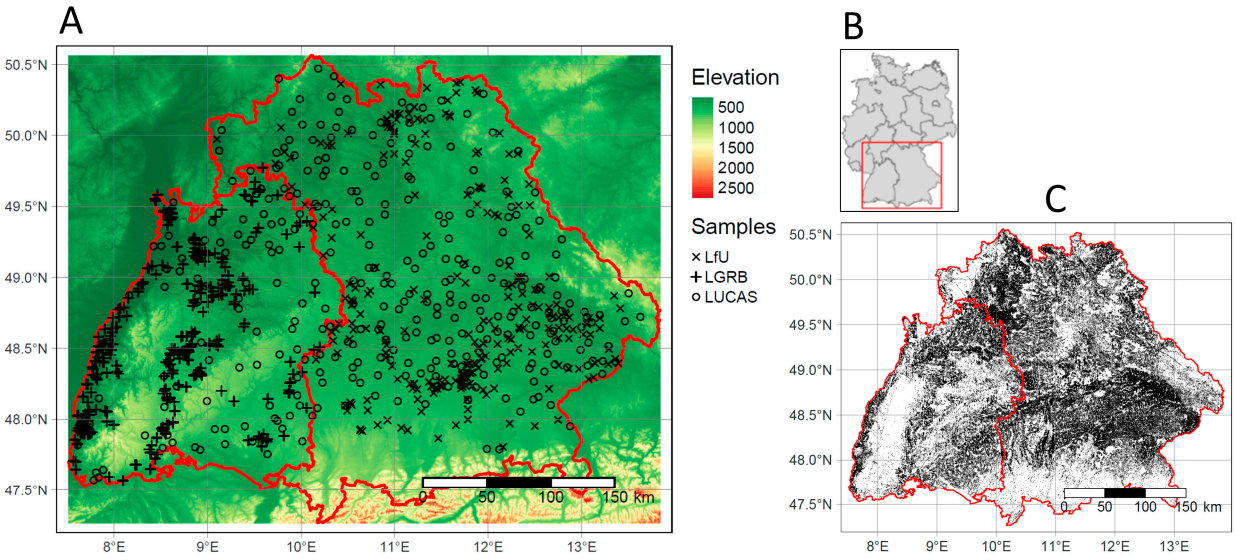

Figure 1. Study area: (**A**) Sampling locations in Bavaria (east) and Baden-Wuerttemberg (west). (LUCAS = The Land Use and Cover Area Frame Survey, LfU = Bavarian Environment Agency, LGRB = State Authority for Geology, Resources and Mining; (**B**) Location within Germany; (**C**) Mask of the exposed soils (black) in the soil reflectance composite.

Table 1. Comparison of both federal states.

	Bavaria	**Baden-Wuerttemberg**
Area (km^2)	71,000	36,000
Area of cropland soils (km^2)	34,000	11,000
Proportion of cropland to total area (%)	48	31
Mean temperature (°C)	8.7	10
Mean precipitations (mm)	836	818
Predominant soil type	Cambisol	Luvisol

2.2. Procedures

The DSM framework is based on the idea that unknown soil properties can be represented by environmental factors (covariates). The covariate layers should be selected to meet all soil forming factors (e.g., climate and topography) and to represent the complex processes behind the soil attribute of interest. Then, a quantitative relationship between measured soil samples and relevant covariates is used to predict values over large areas and to create soil maps. In the current study, we followed the DSM approach for spatial prediction of SOC in the two states of Germany. The procedure included (i) acquiring the

SOC data, (ii) calculating covariates, (iii) model training, and (iv) SOC prediction. The four main components of DSM are discussed in the following sections.

2.3. Soil Samples

The SOC content of cropland topsoil (0–30 cm) was used to train and evaluate the DSM models. A total of 950 soil samples (475 per state) from three different sources were included (Table 2; Figure 1A). The Land Use and Cover Area Frame Survey (LUCAS) was started by the Statistical Office of the EU in 2006 and sampled the soils of over 20,000 individual locations on EU ground [36]. Two sampling campaigns were carried out in 2009 and 2015 on a subset of a 2 km × 2 km regular grid. The LUCAS 2015 dataset includes 837 samples of cropland soils in Germany. If a sampling location was visited in 2009 and 2015, only the more recent measurements were considered. Mixed samples were taken from the topsoil of each location and the SOC was determined using dry combustion. In addition to the LUCAS dataset, measurements from soil legacy databases of the Bavarian Environment Agency (LfU) and the State Authority for Geology, Resources and Mining (LGRB) were included. These point data are allocated to different sampling campaigns (mapping activities, long-term field experiments, etc.) and therefore do not follow a consistent sampling design (Figure 1A). Only samples taken within the analysis period of the exposed soil composite (Section 2.4.1) were considered (1984–2014). The databases contain descriptive information about the soil profiles, which were raised for general soil surveys, as well as further analytic information about the soil properties. As in the LUCAS surveys, the SOC was also analyzed using dry combustion. Only topsoil samples from cropland profiles were considered. This was verified using the mask of the exposed soil composite (Figure 1C). Duplicate measurements from the same sampling locations were removed. All LUCAS samples within the soil reflectance composite were used for the models. As the number of samples in BW was limited, all available samples from the soil database were added to the training data (Table 2). The general availability of soil samples was higher in BY. This is true for samples of the LUCAS campaigns and the legacy data. The number of soil samples in BY was adjusted to the BW dataset with a random selection of samples from the legacy data (Table 2) so that both states can be trained with the same sample size (475). This step was included to increase the comparability of the results and to simplify the sample exchange in mixed-data models (Section 2.8).

Table 2. Summary statistics of SOC measurements for each data source and both states.

State	Source	Samples	SOC (%) Min	SOC (%) Max	SOC (%) Mean	SOC (%) SD	SOC (%) IQR
BY	LUCAS	227	0.6	14.81	2.1	1.68	1.15
	LfU	248	0.54	15.6	3.13	2.8	1.91
	Total	475	0.54	15.6	2.63	2.38	1.54
BW	LUCAS	91	0.79	5.78	1.78	0.87	0.99
	LGRB	384	0.45	8.31	1.48	0.93	0.88
	Total	475	0.45	8.31	1.74	0.92	0.9

2.4. Covariates

The covariates were selected to meet the criteria of the SCORPAN factors [6]. An overview of the covariate groups and layers used in the models is provided in Table 3. The layers were resampled to a common grid cell size of 30 m (1″) using a cubic spline interpolation in GDAL [37]. It has been shown that distance-based covariates can be included in RF models to account for the spatial autocorrelation of the soil samples [8,38,39]. This was done using seven Euclidean distance fields (EDF), as proposed by Behrens et al., (2018) [38]: one for X and Y coordinates, each corner and the center of the research area.

2.4.1. Earth Observation

The Soil Composite Mapping Processor (SCMaP) was developed to automatically analyze multispectral data over a defined period and detect exposed soils to create per-pixel composites for area-wide mapping [14]. A detailed description of the method and processing steps of the SCMaP workflow is provided in the original papers [14,40]. For the bare soil reflectance composite of Germany, all available Landsat 4, 5, and 7 scenes from 1984–2014 were processed. The period of 30 years was analyzed for several reasons. First, it allows to include of a higher percentage of cropland areas, as some soils are rarely exposed. The larger area also increases the number of available soil samples that can be used for modeling. Secondly, a longer period also increases the number of bare soil observations and therefore reduces seasonal effects like moisture or crop residues. A modified vegetation index (PV) was used to distinguish between exposed soils (i.e., croplands) and other land cover classes, like urban areas, grassland, or forests:

$$\mathrm{PV} = ((\mathrm{NIR} - \mathrm{RED})/(\mathrm{NIR} + \mathrm{RED})) + ((\mathrm{NIR} - \mathrm{BLUE})/(\mathrm{NIR} + \mathrm{BLUE})) \quad (1)$$

Thresholds of maximum (PVmax) and minimum (PVmin) index composites were derived to create an exposed soil mask [40] (Figure 1C). The reflectance of the detected exposed soil pixels was averaged per band to create the soil reflectance composites. The used composite map consists of six spectral bands from 0.45 to 2.35 µm with a spatial resolution of 30 m. A description of the bands is provided in Table 3. In the following, the six Landsat bands are referred to as scmap.1–6. The generated exposed soil mask is provided in Figure 1C.

2.4.2. Terrain Attributes

All terrain-based covariates were extracted from the ALOS World 3D-30m DEM, provided by the Earth Observation Research Center of the Japan Aerospace Exploration Agency [41]. The System for Automated Geoscientific Analyses (SAGA-GIS) [42] was used to calculate five terrain-related covariates: topographic wetness index (TWI); valley depth (vdepth); multiresolution index of valley bottom flatness (VBF); and negative and positive topographic openness (openn and openp), as proposed by Hengl and MacMillan (2019) [43]. The contextual spatial modeling (CSM) framework was applied to enhance the representation of different scales in the modeling process [44]. To do so, the DEM was up-scaled six times (90, 180, 360, 720, and 1440 m) and the terrain-based covariates were recalculated at each step (e.g., dem_twi_90-1440). It has been shown that the explanatory power of the terrain covariates can be improved with this process, as some large-scale patterns are only visible if the resolution of the DEM is reduced by aggregation.

2.4.3. Climate Data

Climate data were downloaded from the Open-Data-Server of the German Meteorological Service (DWD) with a spatial resolution of 1 × 1 km [45]. The rasters are based on interpolated mean values of weather and soil measurements at climate stations in Germany. Detailed descriptions of the methods are provided online at the Open-Data-Server [45]. Climate parameters are calculated in 30-year intervals and are only updated every 10 years. The period of 1981–2010 was selected, as it is as close as possible to the calculation of the soil reflectance composite (1984–2014). Most climate parameters are available with 12 rasters as monthly averages. To reduce redundancy, a spatial principal component analysis (PCA) was performed for each climate parameter with more than one raster, using the GSIF package [46]. The most important components of each climate parameter with a combined explained variance of over 85% were kept.

Table 3. Overview of all covariates.

Group	Covariate	Original Resolution	Layers (n)	Abbreviation	Source
Satellite	SCMaP Band 1: blue (0.45–0.52 µm)	30 m	1	scmap.1	[14]
	SCMaP Band 2: green (0.52–0.60 µm)	30 m	1	scmap.2	[14]
	SCMaP Band 3: red (0.63–0.69 µm)	30 m	1	scmap.3	[14]
	SCMaP Band 4: NIR (0.77–0.90 µm)	30 m	1	scmap.4	[14]
	SCMaP Band 5: SWIR1 (1.55–1.75 µm)	30 m	1	scmap.5	[14]
	SCMaP Band 6: SWIR2 (2.09–2.35 µm)	30 m	1	scmap.6	[14]
Soil	Soil type	1:200,000	1	soil_type	[35]
	Soil texture	1:200,000	1	soil_texture	[35]
	Sand content	1:200,000	1	soil_texture_sand	[35]
	Silt content	1:200,000	1	soil_texture_silt	[35]
	Clay content	1:200,000	1	soil_texture_clay	[35]
	Parent material	1:5,000,000	1	soil_geology	[47]
	Geomorphographic class	1:1,000,000	1	soil_geomorphology	[48]
Terrain	Digital elevation model	30 m	1	dem_30	[41]
	Topographic wetness index	90–1440 m	5	dem_twi_90-1440	[42]
	Valley depth	90–1440 m	5	dem_vdepth_90-1440	[42]
	Multiresolution index of valley bottom flatness	90–1440 m	5	dem_vbf_90-1440	[42]
	Negative topographic openness	90–1440 m	5	dem_openn_90-1440	[42]
	Positive topographic openness	90–1440 m	5	dem_openp_90-1440	[42]
Climate	Multi-year means of air temperature (2 m)	1000 m	2	DWD_temp	[45]
	Multi-year means of precipitation	1000 m	2	DWD_prec	[45]
	Multi-year soil temperature at 5 cm depth in bare soil	1000 m	2	DWD_soil_temp	[45]
	Multi-year grids of soil moisture under grass and sandy loam	1000 m	1	DWD_soil_moist	[45]
	Multi-year mean of the number of frost days	1000 m	1	DWD_frost_days	[45]
	Multi-year mean of the number of hot days	1000 m	1	DWD_hot_days	[45]
	Multi-annual mean onset/ending of vegetation	1000 m	1	DWD_vegetation	[45]
	Multi-year mean of the annual climatic water balance	1000 m	1	DWD_water_balance	[45]
	Multi-year mean of the monthly drought index	1000 m	2	DWD_drought	[45]
	Multi-year mean of sunshine duration	1000 m	2	DWD_sunshine	[45]

2.4.4. Legacy Soil Maps

Legacy maps and databases of soils and geological properties are provided by the German Federal Agency of Geoscience and Resources (BGR). The most recent soil map covering Germany is based on a 1:200,000 scale (BÜK200) and is composed of 55 individual tiles that have been harmonized into a common format [35]. The vector format map is based on expert knowledge and the German soil mapping standards KA4/KA5 [49]. The polygons are connected to a database containing details about soil types and sub-types as well as information about the properties of the individual soil horizons. The main soil types of the German Soil System were extracted and transformed into a raster layer. It was verified that both states share the same soil classes, as this is necessary for the models to be transferable to the other state. For illustration, the soil types were converted into the system of the World Reference Base (Table 4). Carbon-rich bog soils were excluded from the modeling area with the BÜK200 to avoid biases arising from outliers with very high SOC contents. Information on topsoil texture was extracted from the first mineral horizon. Separate layers were created for the corresponding sand, silt, and clay shares using mean values of the soil texture classes [47]. The Geological Map of Germany (1:5,000,000) [44] and the Map of Geomorphographic Regions in Germany (1:1,000,000) [45] were used to extract information about geological properties.

Table 4. Soil types and mean SOC content at the sampling locations. Main soil types were extracted from the German Soil Map (BÜK200) [35] and converted to the units of the World Reference Data Base.

Soil Type		BY		BW	
German Soil System	**World Reference Base**	**Samples**	**Mean SOC (%)**	**Samples**	**Mean SOC (%)**
Vega	Fluvisols	26	2.69	66	1.67
Braunerde	Cambisols	217	1.64	70	2.1
Pelosol	Vertisols	26	1.96	31	2.72
Gley	Gleysols	53	3.44	22	2.66
Anmoorgley	Gleysols	37	8.76	1	3.57
Parabraunerde	Luvisols	35	1.77	201	1.37
Rendzina	Leptosols	14	3.14	24	2.23
Pararendzina	Regosols	44	2.77	52	1.64
Pseudogley	Planosols	15	2.02	1	1.16
Kolluvisol	Colluvic	8	1.34	7	1.47

2.5. Similarity Analysis

The spatial variation of the natural conditions was analyzed to evaluate if both states are similar enough to successfully transfer the DSM models. Multivariate Environmental Similarity Surfaces (MESS) were created using Maxent [50]. The MESS is calculated by comparing the covariate values at the training locations with the values of the research area. The resulting map can be used to evaluate the area of applicability of the models and to distinguish regions where covariates are outside the range of the training data. Two uncertainty maps were created for the whole research area, one with the training data of each state. Based on the feature selection, the most important climate, terrain, and soil covariates were used to calculate the similarity of the natural conditions. This approach is similar to the study by Meyer and Pebesma (2021) [51] but does not assign weights according to the feature importance. Additionally, density plots were created to compare the values of the most important spectral bands at the sampling locations of both states. This process was repeated for other important numeric covariates from the climate and terrain group.

2.6. Random Forest

RF models were used as spatial prediction models. The algorithm creates a large array of decision trees that are trained with a random subsample of the input data and use random features at each decision node [52]. This approach is widely used because RF models can easily deal with a high number of covariate layers and achieve good prediction results with a minimal amount of hyperparameter tuning [8,53]. The models were trained and evaluated with the R package h2o (version 3.38.0.1) [54]. To improve comparably, all RF models were trained with the same default set of parameters (ntree = 200, mtries = number of predictors divided by 3). Categorical variables were preprocessed with one-hot encoding. A variable importance ranking was conducted with the h2o package [54] to identify the most important covariates for the combined models in both states. The importance of every feature is estimated by calculating the attributed reduction of error (i.e., the variance reduction of the response value) at each decision node. To reduce the risk of overfitting, the training of combined models was performed in two steps. First, the variable importance was calculated for all covariates. Subsequently, covariates with relative importance above 1% in either state were selected and used in the final combined models in both states.

2.7. Accuracy Assessment

The prediction accuracies of RF models were assessed based on the results of a 10-fold cross-validation. For each of the five covariate groups, the process was repeated six times. R^2, root-mean-square error (RMSE), mean absolute error (MAE), Lin's concordance correlation coefficient (CCC), as well as the out-of-bag error (OOB) were calculated from the results of each cross-validation run. The differences between the groups were tested using an

ANOVA and Tukey's honest significance test (HSD). Plots of the results were created with the package ggpubr [55] and the significance levels were illustrated with stars: * $p < 0.5$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$. To check if the models under- or overestimated the SOC values, a locally estimated scatterplot smoothing (LOESS) function was applied to the cross-validation results. The mean R^2 of the cross-validation results was calculated to compare the results of the mixed-data models and the control models. This was done for each of the five covariate groups and both states. Differences between both groups were assessed with standard error (SE) of the means.

2.8. Transferability of Different Covariates

To test the transferability of different covariates, five groups were created based on the primary data sources: (i) remote sensing data (satellite), (ii) terrain attributes (terrain), (iii) climate data (climate), (iv) soil legacy maps (soil), and (v) a combination of the most important covariates from all groups (combined). An overview of the workflow is shown in Figure 2. Soil samples from both states and environmental covariates from different sources were harmonized into a common raster format. In the first step, models for BY and BW were trained and assessed separately. Each of the five covariate groups was used to predict the SOC content of the respective state. The results were used as a baseline to assess the prediction accuracy of each covariate group before the models were transferred. The transferability was evaluated in two steps. First, the pre-trained models of the previous step were tested with the soil samples of the respective other states (Figure 2A). It is not uncommon that models are extended to regions with low or no availability of soil data. To simulate this condition, soil samples from both states were used to train the models (mixed-data models) while only soil samples from one state were used to test the model accuracies. For each state and covariate group, nine proportions were tested (10–90%) (Figure 2B). The percentages correspond to the number of training samples from the same state as the test data. Control models were set up with the same training samples but without the additional data from the other state (Figure 2C). The training data of both states were divided into ten equal parts using stratified random sampling. The random strata were then assigned to both groups manually to ensure that the samples are identical and that no bias is introduced with the training data. Differences in the accuracy of both groups can thus only be explained by the additional training data from outside the validation area.

2.9. Prediction of the SOC Maps

A final map of the SOC in both states was predicted using the RF models with the highest accuracies and the ranger package [56]. SOC values were predicted for BY and BW separately and subsequently combined into a final raster. To illustrate the influence of the satellite and soil covariates on the prediction results, additional maps were generated using the individual covariate groups. The results are presented in magnified sites to showcase small-scale variations. In the last step, the satellite models of both states were transferred and used to predict the SOC of the respective other state. To examine the effectiveness of model transferability, two sites were selected in each state: one in an area with high similarity (i.e., the soil forming factors explaining variation in SOC are quite similar in the two states) and one in an area with low similarity (i.e., the soil forming factors explaining variation in SOC are different in the two states). The results were then compared to the SOC predictions of the non-transferred satellite models.

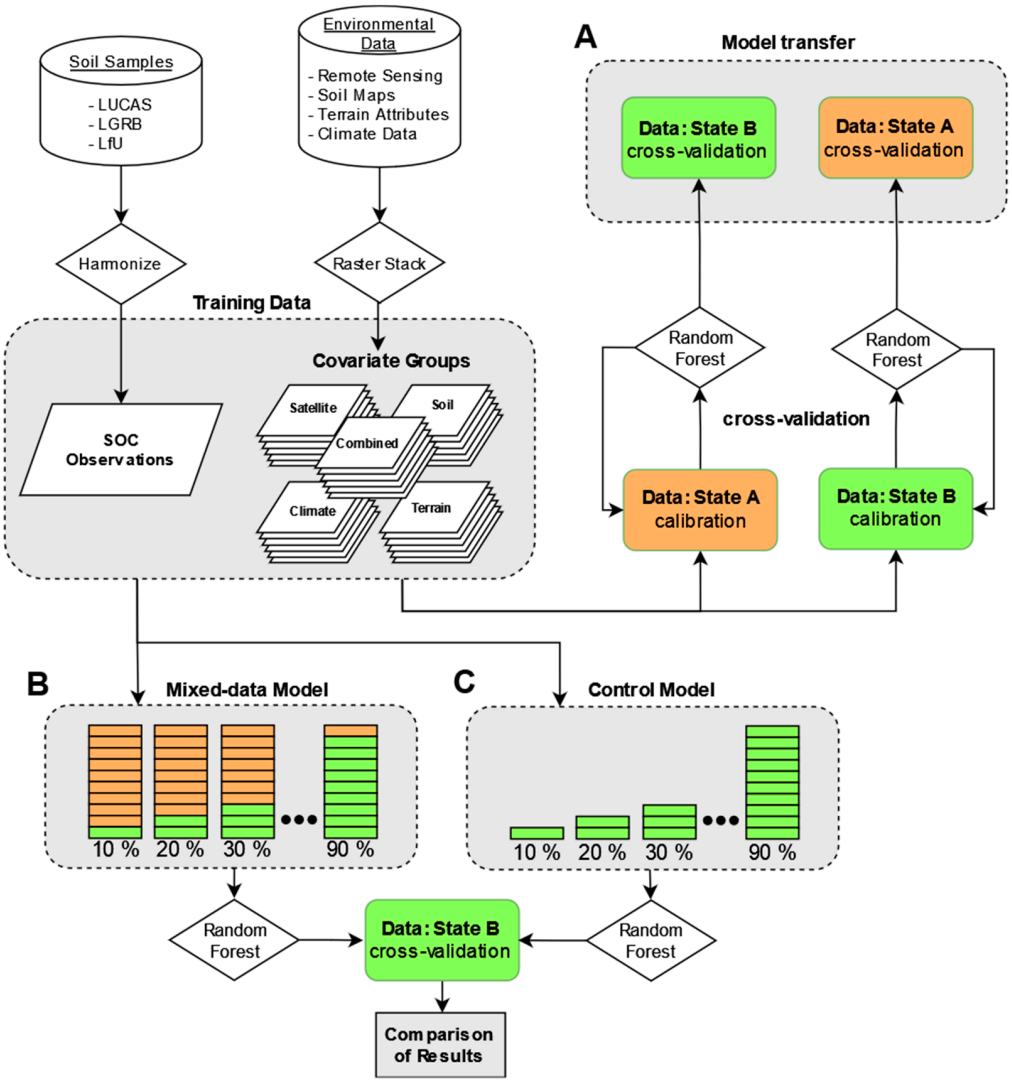

Figure 2. Flowchart of the SOC prediction and model transfer. (**A**) For each covariate group, models were trained with soil samples from one state and tested with samples from the other one. (**B**) Mixed-data models were trained with samples from both states to evaluate the effect of additional data. Nine models (10–90%) with different proportions were trained for all covariate groups and both states. The results were compared to control models built without the additional data (**C**).

3. Results

3.1. Performance of the Baseline Models

The cross-validation results for each covariate group and the combined models are provided in Table 5. A median RMSE between 1.37 and 1.6% of SOC was calculated. Combined models show significantly higher accuracies than all other covariate groups and produced the best results in both cases (R^2: BY = 0.68, BW = 0.48; RMSE: BY = 1.42%, BW = 1.37%). The results of the other covariate groups were different in both states (see Figure 3). Satellite (R^2 = 0.53) and especially soil models (R^2 = 0.61) produced good results in BY. In contrast to that, terrain models produced the second-best results in BW (R^2 = 0.35) while satellite models showed the overall lowest accuracy (R^2 = 0.30). In general, R^2 and RMSE were lower in BW.

Table 5. Cross-validation results (R^2, RMSE, CCC, MAE, and OOB) of the baseline and transferred models in both states.

		Baseline Models					Transferred Models			
	Group	R^2	**RMSE**	**CCC**	**MAE**	**OOB**	R^2	**RMSE**	**CCC**	**MAE**
BY	combined	0.68	1.42	0.81	0.74	1.93	0.34	1.68	0.47	0.57
	satellite	0.53	1.53	0.71	0.88	2.69	0.42	1.63	0.5	0.58
	soil	0.61	1.48	0.77	0.79	1.97	0	1.91	0	0.76
	terrain	0.44	1.58	0.63	0.96	3.18	0	2.01	0	0.87
	climate	0.42	1.6	0.61	0.92	3.31	0	1.99	0	1.2
BW	combined	0.48	1.37	0.63	0.44	0.4	0.36	1.43	0.44	0.91
	satellite	0.3	1.44	0.48	0.52	0.53	0.25	1.50	0.41	0.87
	soil	0.31	1.43	0.5	0.49	0.51	0	1.64	0	1.15
	terrain	0.35	1.41	0.54	0.51	0.50	0	1.60	0	1.11
	climate	0.31	1.43	0.5	0.54	0.52	0	1.62	0	1.22

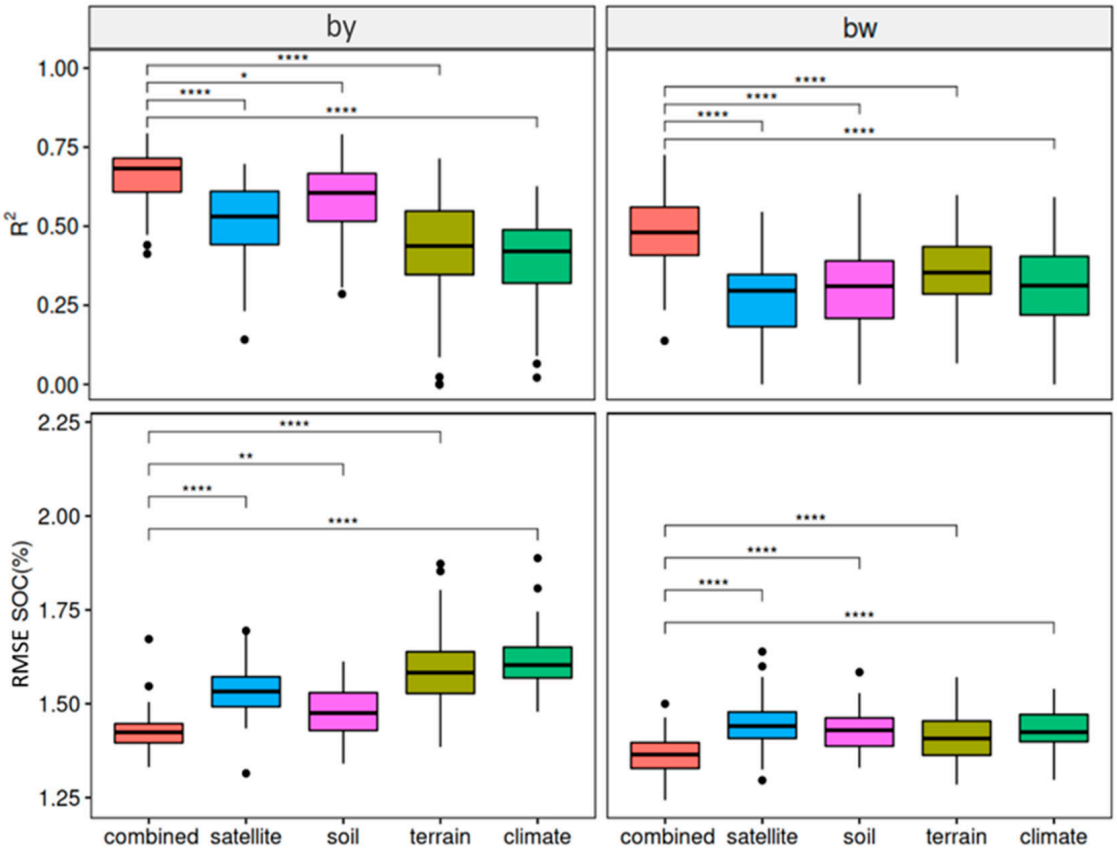

Figure 3. Boxplots (R^2 and RMSE) of the baseline models. Significant differences between combined models and the other covariate groups are indicated with stars (* $p < 0.5$; ** $p < 0.01$; **** $p < 0.0001$).

Scatter plots of predicted and measured SOC values are illustrated in Figure 4. The models in BW showed a clear tendency to overestimate low and underestimate high values. Some models in BY had the same issue but a general underestimation of middle to high values seems to be apparent in most models. Soil and combined models produced the best results for high SOC values in BY and therefore also the overall highest R^2 values.

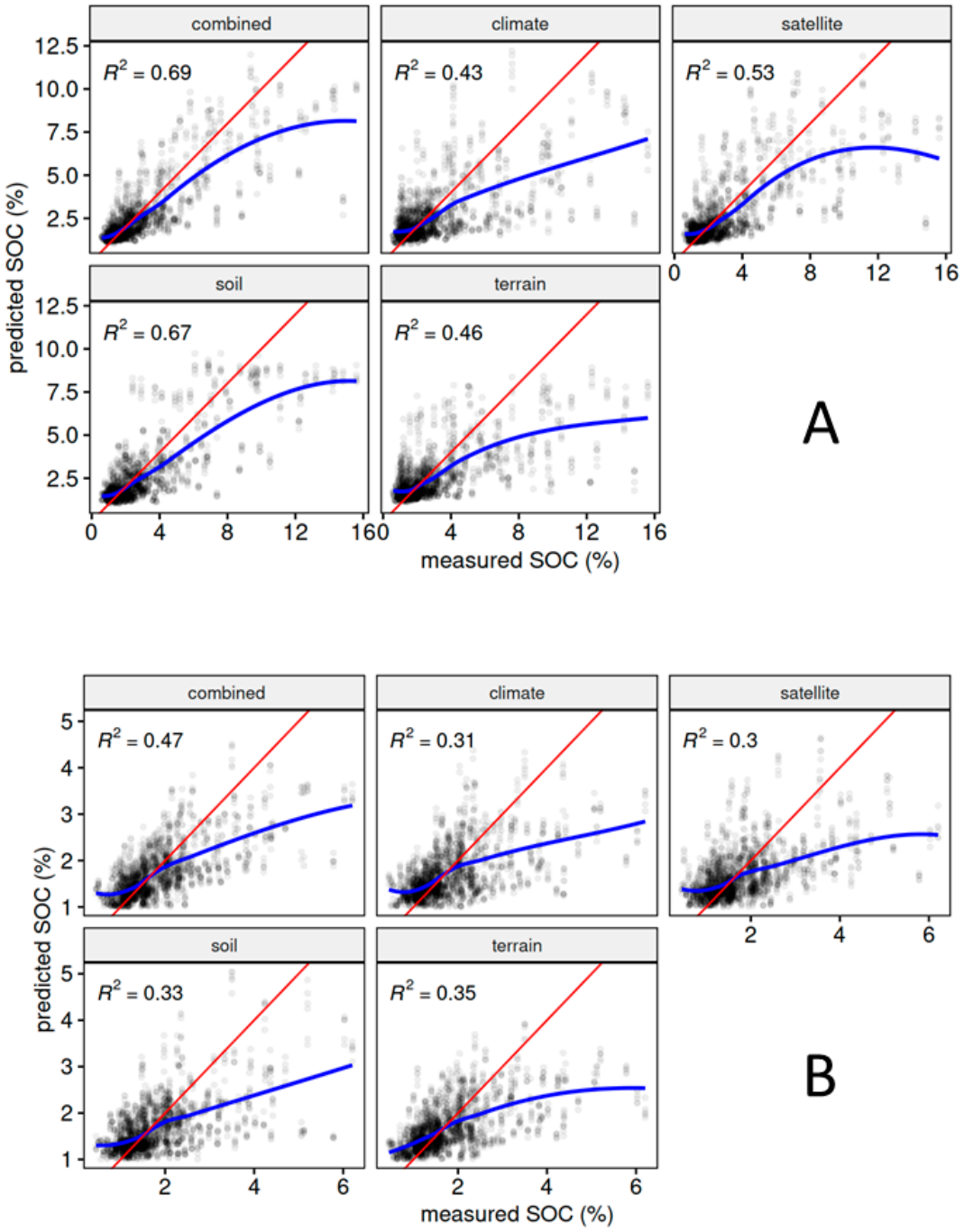

Figure 4. Scatter plot and LOESS-function (blue) of the predicted and measured SOC values in BY (**A**) and BW (**B**). R^2 values were calculated based on all cross-validation results (6 times 10). 1:1 lines are displayed in red.

The results of the variable importance ranking of both states are shown in Figure 5. Spectral bands of the exposed soil composite were highly important for the models in both states. The SCMaP band 1–4 contributed 55% to the combined models of BY and 26% of BW. The categorical covariates for soil type, geology, and geomorphology also showed a high importance in both cases. In general, fewer covariates were classified with importance higher than 1% in BY. This is especially true for terrain covariates which were more important for models in BW.

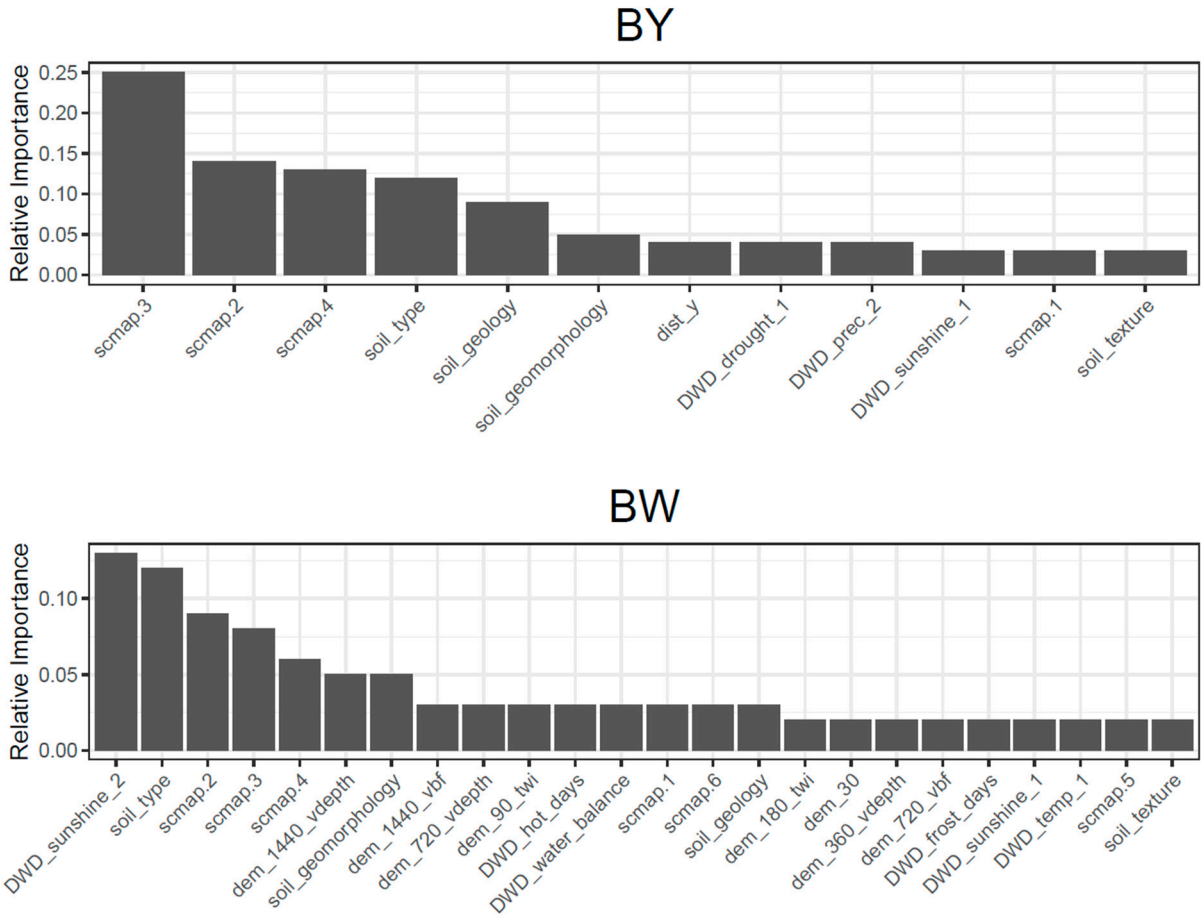

Figure 5. Relative variable importance of covariates in the combined models. Only covariates with relative importance above 1% are shown.

3.2. Performance of the Transferred Models

The results of the transferred models are presented in Table 5. Considering the comparatively low model performances for the soil, terrain, and climate group, it is clear that only combined and satellite models were successfully extended to the other state. Satellite models produced the best results in BY (RMSE = 1.63%) while the combined model showed better results in BW (RMSE = 1.43%). Differences between combined and satellite models before and after the model transfer are displayed in Figure 6. As expected, the R^2 declined in all cases but in different magnitudes. Generally, the decline was more significant for the models in BY and combined models in both cases.

3.3. Performance of the Mixed Data Models

The results of the mixed-data models are shown in Figure 7. R^2 values are plotted against the results of the control models (blue line) to evaluate the effect of the additional data. The X-axis corresponds to the percentage of training data from the test area. It can be seen that the R^2 values at 90% are similar to the results of the baseline models while the results at 10% are similar to the results of the transferred models (Table 5). The R^2 of the control models increased with the number of training samples for each covariate group. This trend was more pronounced for small training sets (10–40%) and saturated with increasing size (e.g., Figure 7C,D).

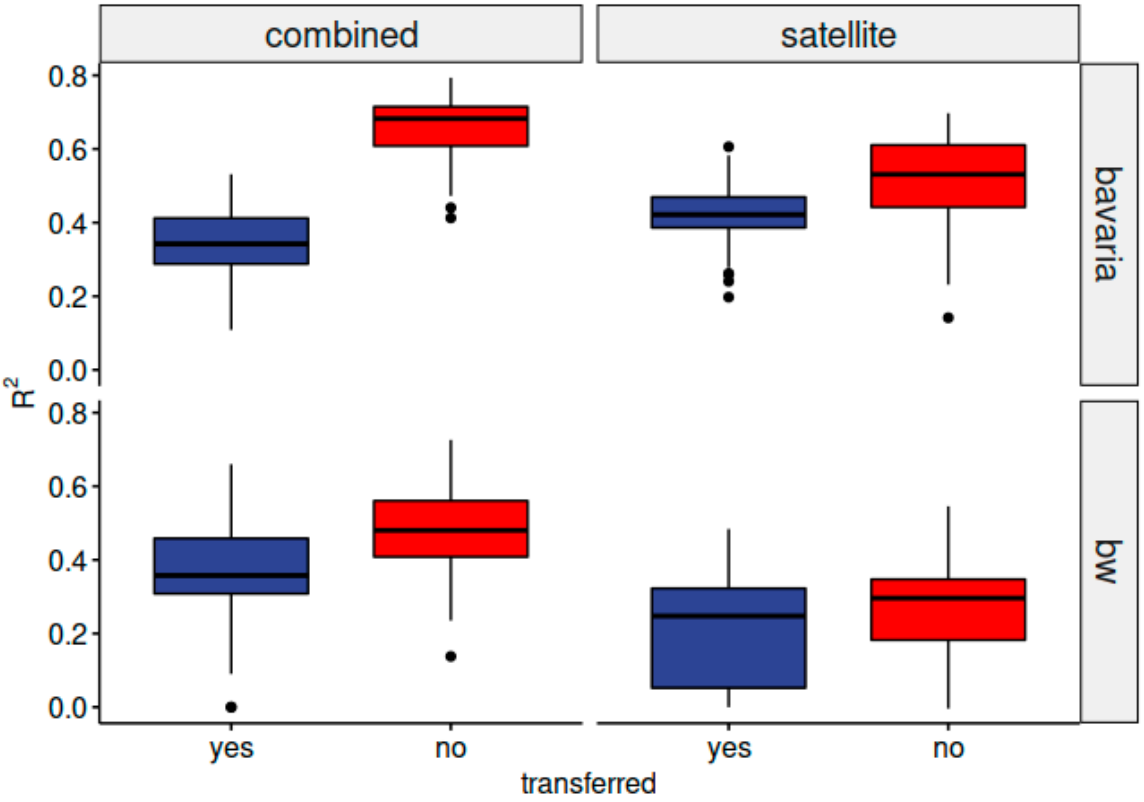

Figure 6. Comparison (R^2) of combined and satellite models before (red) and after (blue) the models were transferred and tested with the samples of the other state. Black dots indicate outliers.

Additional training samples in the mixed-data models improved the accuracies of combined, satellite, and terrain models, as illustrated by the SE. This trend is more pronounced when the proportion of training data from the test state is low (10–20%) and is illustrated by the fact that the SE of control and mixed-data models is not overlapping. The improvement is greatest for satellite models in both states (Figure 7C,D) and combined models in BW (Figure 7B). In contrast to all other groups, mixed-data models decreased to accuracies of soil models in both states (Figure 7E,F). No effect was observed for the climate models (Figure 7I,J).

3.4. Final Prediction Maps

The final SOC map, based on the predictions of the combined models, is presented in Figure 8. It is visible that the highest SOC values generally occur in southern BY and are connected to the main streams, originating from the Alps. On the other hand, low SOC values are present in the north and west of BW, as well as in between the carbon-rich streams in southern BY. Magnified scenes of the predictions based on different covariates are presented in Figure 9. The maps illustrate local variations of the SOC as well as the influence of different covariates on the prediction results. In the case of BW, the local patterns of the SOC were similar when predicted with satellite (Figure 9D) or combined models (Figure 9F). This is not the case for BY, where clear differences between both models are visible (Figure 9C,E). In contrast to the other maps, the outlines of the polygons are visible in the SOC maps of the soil models (Figure 9A,B). However, this mosaicking effect is barely visible in the SOC maps of the combined models, even though the soil covariates were included. The maps in Figure 10 showcase the effect of the transferred satellite models on the local SOC predictions. Sites with dissimilar soil conditions between the states are compared to sites with similar soil conditions. It is visible that satellite models trained in BY overpredicted high SOC values in BW (Figure 10C,D) while the models trained in BW

underpredicted the values in BY (Figure 10A,B). This effect is less pronounced for the sites with similar soil conditions (Figure 10B,D).

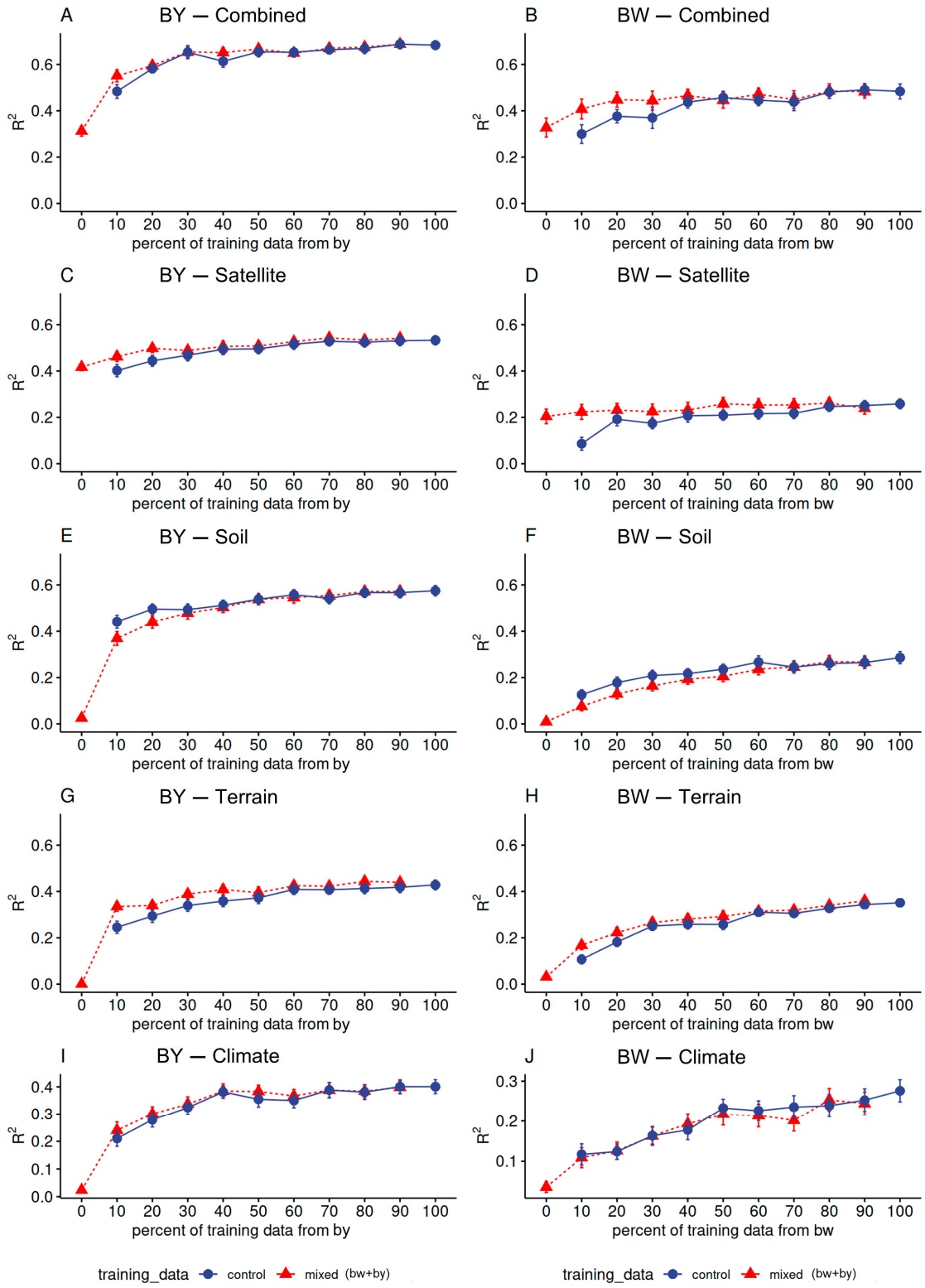

Figure 7. Results (R^2) of mixed-data (red) and control models (blue) for each covariate group. Mean values and SE of the 6 times 10-fold cross-validation are shown. The x-axis corresponds to the percentage of training samples from the state in which the models were tested. Control models were trained with the same data but without additional samples from the other state.

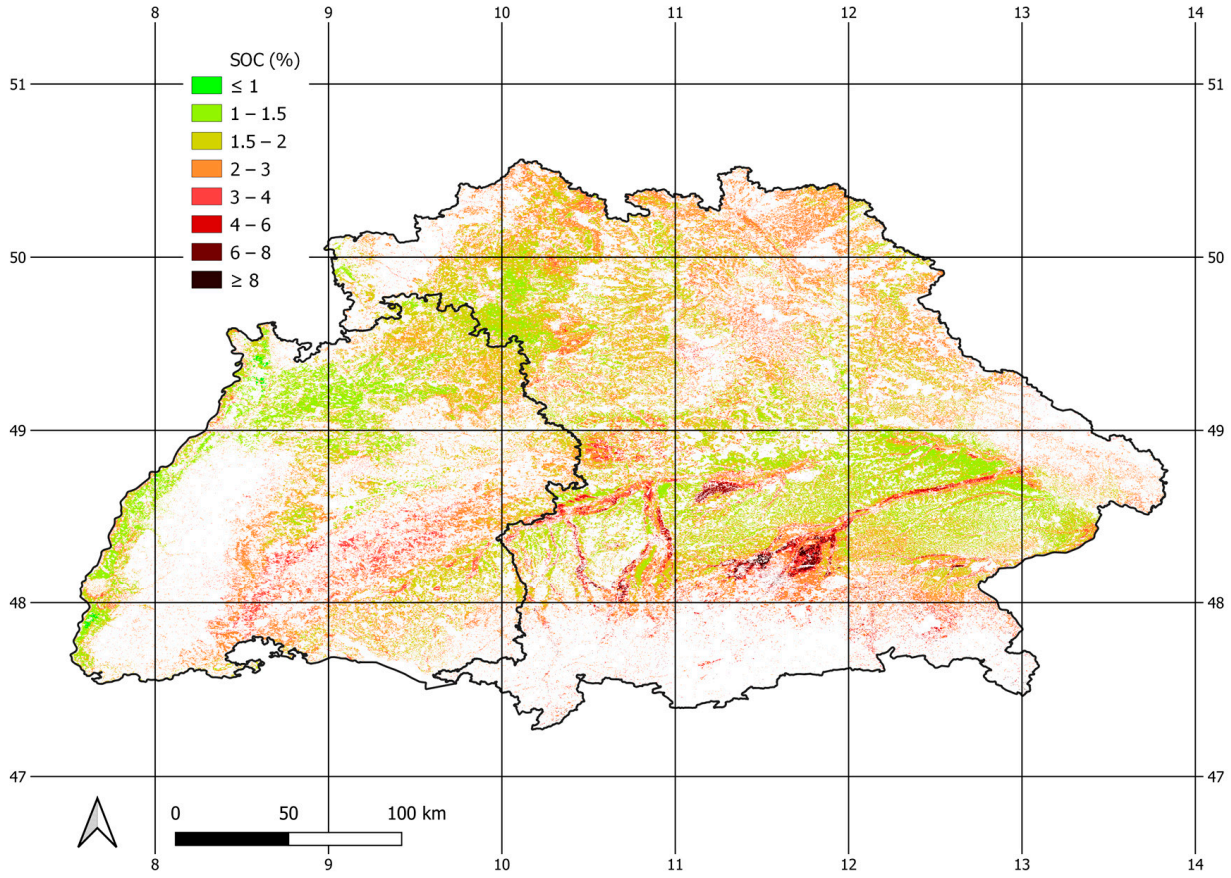

Figure 8. Final SOC map. Values were predicted for BY and BW independently, using the combined models of the respective state.

3.5. Similarity of the States

The MESS was calculated for both states using the locations of the training samples and the most important covariates (Figure 11). Negative values (red areas) correspond to covariate values that are not represented in the training data. The results show that the mountain ranges are not represented in the training data of either state. This was expected as no soil samples of these regions were included and almost no exposed soils were detected (Figure 1C). Training samples of BW represent most parts of central BY but show low similarity with the most northern parts and some regions in the east (Figure 11A). In turn, the training samples of BY are most similar to regions close to the border of BW but do not represent the covariate values of the Rhine Valley, in the west (Figure 11B).

Density plots of important satellite, climate, and terrain covariates at the sampling locations are shown in Figure 12. The spectral bands are mostly normally distributed and the peaks are congruent. This is not the case for most of the other analyzed covariates. Especially, the distributions of the terrain covariates are fundamentally different between the states.

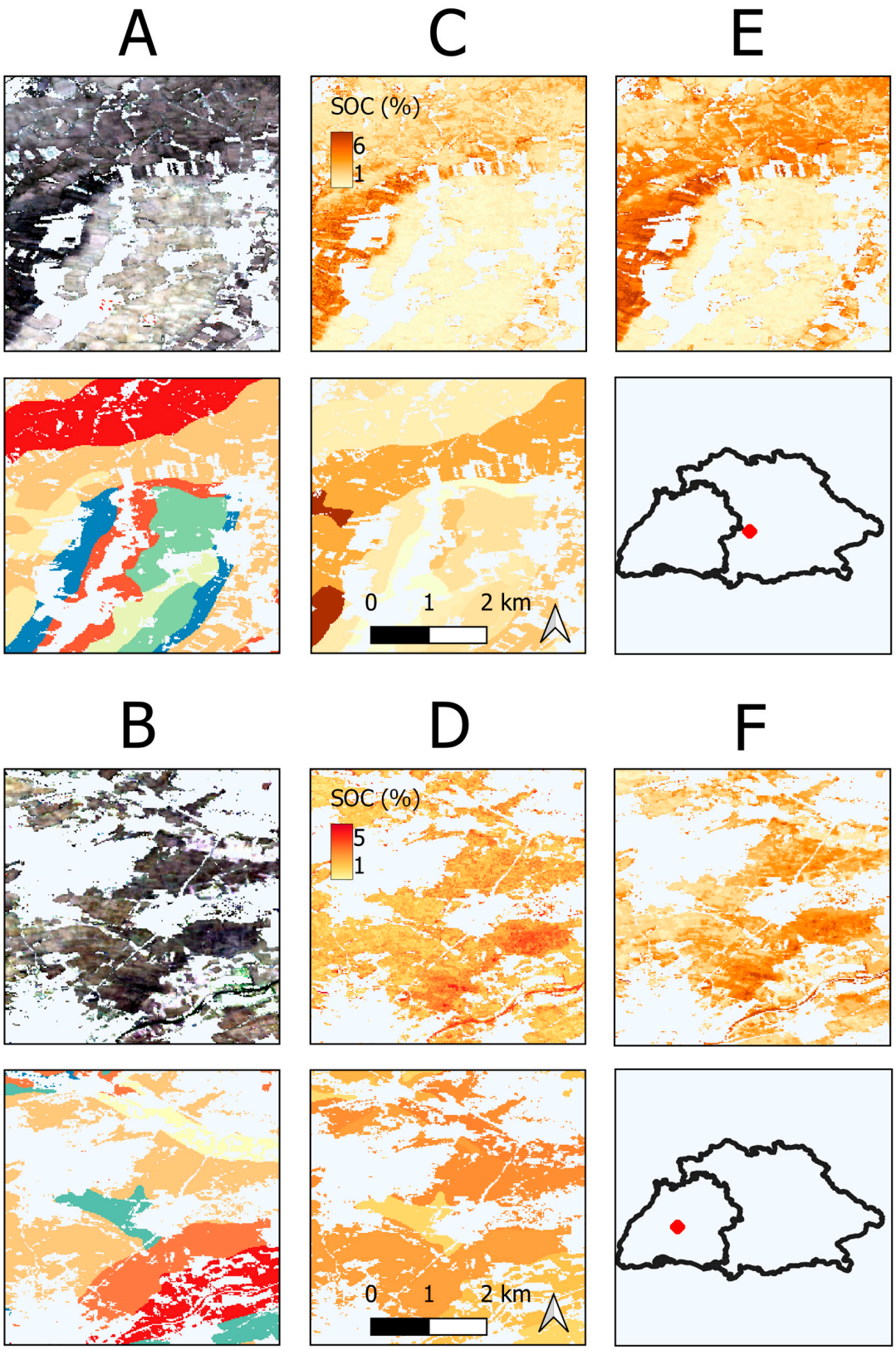

Figure 9. Comparison of the results from different models in BY (**top**) and BW (**bottom**). Examples of the exposed soil composite and soil types as input covariates (**A**,**B**); Respective SOC prediction for satellite and soil models (**C**,**D**); SOC predictions of the combined models (**E**,**F**). Red dots show the locations within the research area.

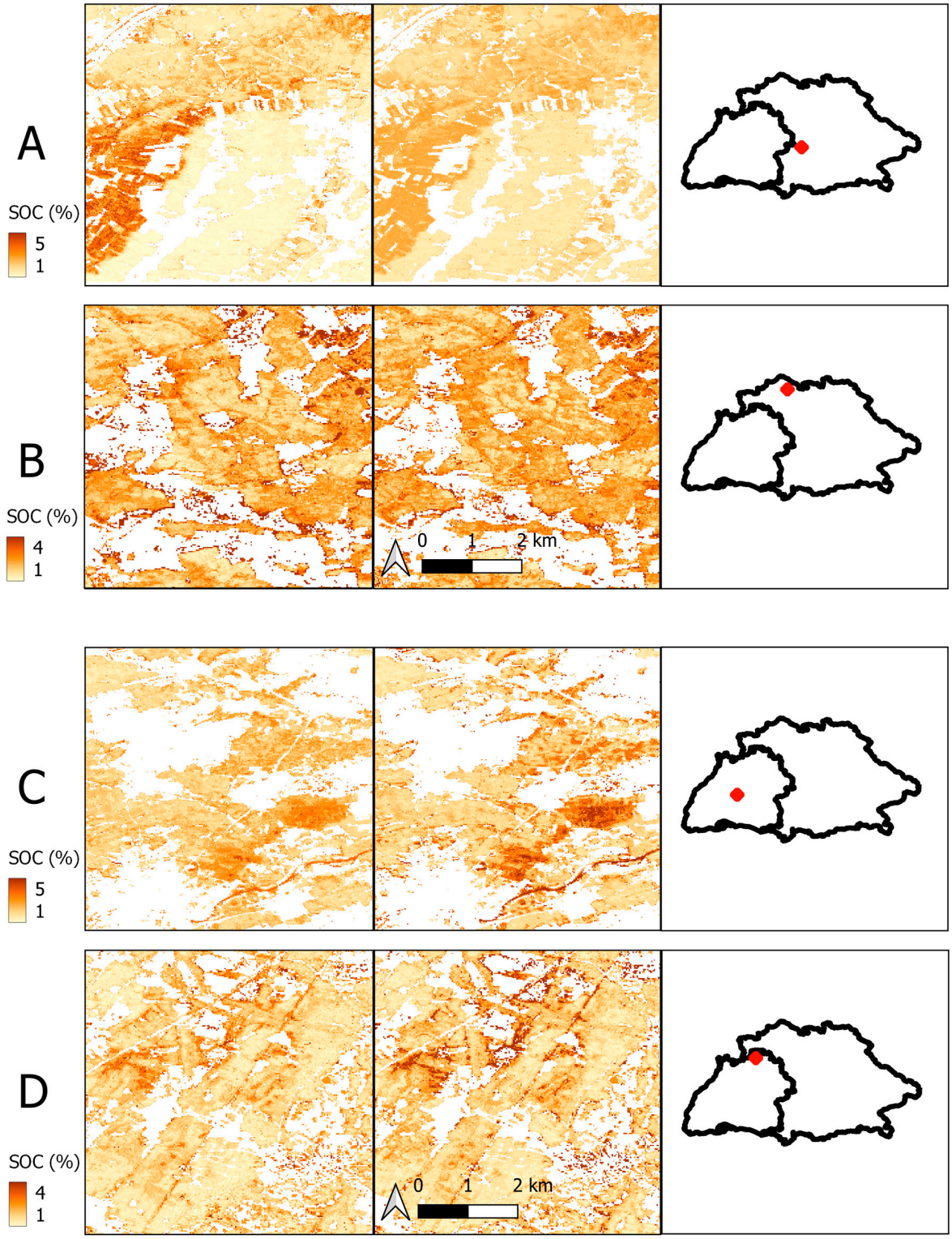

Figure 10. Influence of the model transfer on the SOC prediction for satellite models in BY (**A**,**B**) and BW (**C**,**D**). Maps of non-transferred models (**left**) and transferred models (**right**). Sites with dissimilar soil conditions in both states (**A**,**C**) are compared to sites with similar soil conditions (**B**,**D**). Red dots show the locations within the research area.

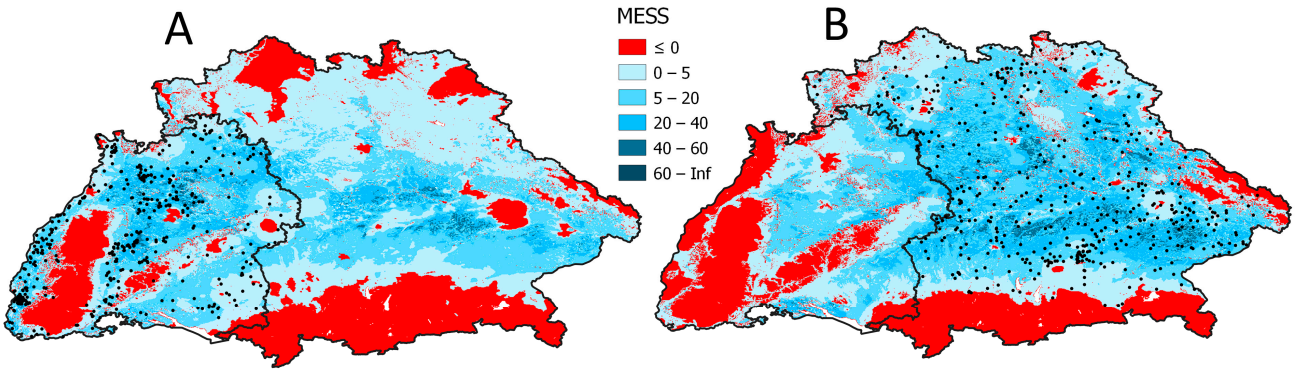

Figure 11. Multivariate Environmental Similarity Surfaces (MESS), calculated with the most important covariates and training samples of BW (**A**) and BY (**B**). Areas with high model uncertainty are colored in red. Black dots show sampling locations.

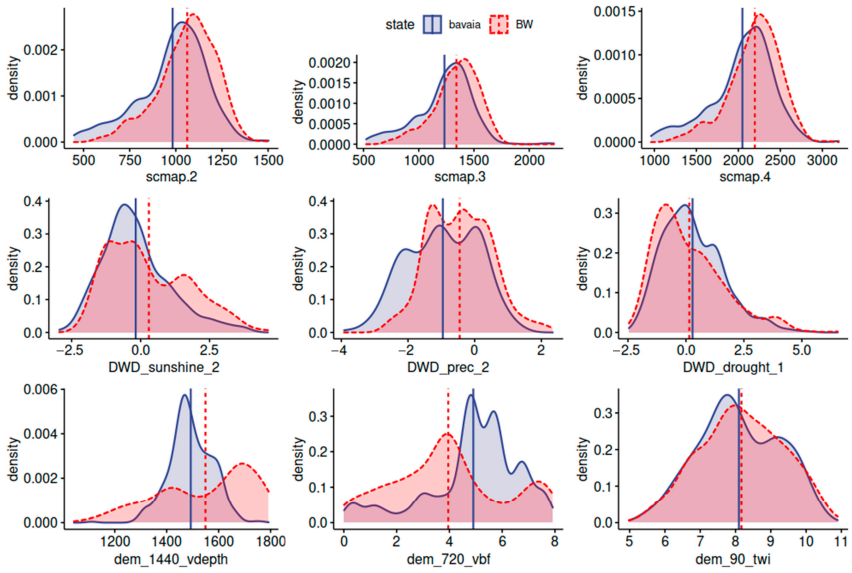

Figure 12. Density plots and mean values (vertical line) of important satellite, climate, and terrain covariates at the sampling locations of both states.

4. Discussion

4.1. General Model Performance

We compared DSM models of SOC in two German federal states to assess the transferability of five different covariate groups. To set a baseline, models were first trained and assessed without being transferred. The best results in both states were achieved with combined models that were trained with the most important covariates from all groups (see Table 5) (R^2: BY = 0.68, BW = 0.48). This finding is coherent with the concept of DSM and the idea to include a wide range of environmental covariates to represent and predict complex soil properties such as soil type or SOC content. Vaudour et al. (2022) [57] conducted a review to analyze the usage of satellite imagery for the prediction of topsoil SOC. Most models were solely based on soil reflectance composites derived from time-series analyses. The majority of studies were conducted in small- to medium-sized regions and R^2 values between 0.5 and 0.7 were reported [18,19,58,59]. A lower accuracy was achieved in a model on a continental scale (R^2 = 0.35) [17] while R^2 values above 0.8 were reported for observations on field level with Sentinel-2 and high-resolution data [60]. Considering the relatively large size of the research area, combined models produced good results, especially in BY. However, the RMSE in this study is still relatively high (BY = 1.42, BW = 1.37) and the

models should not be used for final SOC predictions without further optimization. As seen in the results of the literature, the error also depends on the size of the research area. Large areas, such as the one in this study, usually cover more diverse landscapes and are therefore more prone to error. Still, the RMSE can be improved by model optimization and hyperparameter tuning. However, this was not the main objective of this study, as the comparability between the different covariate groups was more important to describe the transferability.

Ensemble models based on multiple individual ML algorithms have been successfully used to optimize the prediction of SOC [10] and improve the extrapolation of spatiotemporal DSM models [61]. Mixed results were reported for the relative importance of additional covariates in comparison to the satellite images for the prediction of SOC. In one case, the integration of satellite time-series data did not improve the model accuracy [33]. In other studies, the remote sensing images were the most relevant covariates but the importance also depended on the research area [9,10,62]. The results of our study show that the satellite covariates were most important for the combined models in both states, closely followed by the soil covariates (Figure 6). Comparing the individual covariates groups, soil models performed slightly better than satellite models in both cases (Table 5). This mismatch can be explained by the fact that the satellite models underestimated high SOC values, as illustrated by the LOESS function (Figure 4). This resulted in a reduction in the overall R^2 of the satellite models and is less pronounced in soil models, as carbon-rich gleysols are easily segregated from other soil types. As illustrated by the combined models, satellite covariates were generally more important for the prediction of SOC (Figure 5). However, the soil models showed better results for areas with high SOC concentrations and therefore produced high R^2 values. This demonstrates the fact the accuracy of remote sensing data can be significantly increased by adding further covariates from different sources, especially soil properties from legacy maps.

4.2. Differences between the States

The models in BY generally produced better R^2 values while the RMSE was slightly lower in BW (Figure 3). It has been shown that the R^2 correlates with the SD of the soil samples, which is significantly higher in BY [63] (Table 2). This can be explained by dissimilarities in the natural soil conditions of both states. High SOC values are mainly present in the North Alpine Foreland Basin of BY and are allocated to high soil moisture [64]. Even though bog soils were excluded, carbon-rich Gleysols were more abundant in BY (Table 4). As shown in Figure 4, this fact also influences the prediction results of the models in BY. Another reason for different prediction results between states might stem from differences in soil data collection and analysis protocols. The inclusion of soil samples from legacy databases can introduce uncertainties resulting from different sampling methods and imprecise localization. Sampling design is another important factor of DSM [65]. The proportion of LUCAS samples was higher in BY which led to a better representation of the research area and less spatial clustering (Figure 1) [66]. From the results in Figure 7 (A), it can be seen that only 30% of the samples in BY were necessary to maximize the prediction accuracy of the combined models. This trend is different in BW where the accuracy incrementally increases with sample size (Figure 7). The trend can be explained by the spatial clusters in BW and the fact that more samples are necessary to reach the same representation of the research area as in BY. Furthermore, the redundancy of soil information might be higher in BY, as fewer samples are necessary to reach a stagnating accuracy.

The MESS and the density plots of terrain covariates illustrate significant topographic dissimilarities between both states (Figures 11 and 12). This is especially true for the Rhine Valley in the west of BW which is heavily cultivated and very dissimilar to the sampling locations in BY (Figures 1 and 11). Overall, the terrain covariates were more important for the models in BW (Figure 6). The landscape in BW is diverse and terrain attributes might influence the distribution of SOC to a larger extent than in BY, which is mostly dominated by North Alpine Foreland Basin.

Surprisingly, satellite models had the lowest overall R^2 of all covariate groups in BW (Table 5). This result is not in line with the high accuracies that were reported using soil reflectance composites to predict SOC [18,19]. The differences between satellite models in both states could also be explained by the lower variance of SOC values in BW. As described in Section 2.1, the geologic formations in BW are diverse and affect the soil formations as well as SOC. The multispectral images may not provide enough information to differentiate between the different soil types and detect small variations in SOC content. This argument is supported by the fact that the accuracy of the combined models is significantly higher and that the soil types were identified as important covariates (Figure 5). Effects such as soil moisture or crop residue on the generation of the exposed soil composite could be another reason why it is more difficult to predict SOC values with a lower variance [67,68]. These problems may be improved by the availability of large-scale hyperspectral images [5,69,70], and comprehensive measurement campaigns to improve the representation of cropland soils [71]. Nevertheless, the multispectral bands were still valuable for the prediction of SOC, as seen from the variable importance rating of combined models (Figure 6).

4.3. Transferability of Different Covariate Groups

Satellite imagery has been successfully used to predict the SOC content on many occasions (e.g., [10,19]). However, little effort has been made to use remote sensing data to transfer SOC models to extended research areas. The results of this study show that satellite covariates are highly transferable and can be used to improve model extrapolation (Table 5). This is supported by the fact that the accuracy of the satellite models improved when soil samples from another state were added to the training data (Figure 7C,D). The prediction maps of the transferred satellite models show that the results are similar to the original models (Figure 10). Overall, the results indicate that the transferred models based on the soil reflectance composite can serve as a basis for the prediction of SOC in unknown regions. However, the SOC maps based on the transferred models also illustrate the influence of the input samples. Mean SOC values were generally lower in BW (Table 2) and the transferred models underpredicted the SOC in BY (Figure 10A). Differences in the distribution of the input samples should therefore be considered when transferring satellite models. Subsequently, the results can be improved and fine-tuned with the addition of soil samples from the prediction area, as shown by the results of the mixed data models (Figure 7C,D). It has been shown that soil reflectance, especially in VNIR–SWIR region, is closely correlated to the accumulation of organic matter in the topsoil [72,73]. For this reason, efforts have been made to adapt SOC models that were trained with laboratory soil spectral libraries to remote sensing data [67,74]. The results of this study suggest that a similar approach could be used to improve the transferability of DSM models.

Terrain covariates have been regularly used to transfer DSM models [24–30] and it was unexpected that the transferability was lower in comparison to the satellite data (Table 5). A possible explanation could be the fact that the terrain attributes of both states are not similar enough (Figures 11 and 12). The problem was improved by the additional training data. In this case, the mixed-data models were able to adjust for the dissimilarities of the terrain attributes and improve the prediction accuracy in comparison to the control models (Figure 7G,H). Hence, the results suggest that the terrain covariates are only transferable under optimal conditions and if the extended research area is similar to the training data. This is fundamentally different from satellite covariates and should be considered when terrain covariates are used to transfer DSM models.

Models based on soil covariates showed no sign of transferability. One possible explanation could be the natural dissimilarities of soil conditions in both states. This is illustrated by the imbalance of the soil classes in both states, as summarized in Table 4. All soil classes must be present in both states for the models to be transferable. However, because of different natural conditions, some soil types are underrepresented in samples of BW. This can introduce biases during the model training which are then transferred to the other state and lead to poor results. In contrast to terrain and combined models, the

accuracy of mixed-data models dropped in comparison to the control models (Figure 7E,F). These results suggest that the soil covariates were not transferable and should therefore only be used in a local context. Unlike EO data, underlying soil legacy maps are based on various soil mapping standards and extrapolated measurements. Similar conclusions were drawn in the study of Zhao et al. (2020) who only used information from soil legacy maps to correct the results after the models were transferred to an extended area [28]. To improve the transferability of soil covariates, the soil condition needs to be as similar as possible to reduce the errors resulting from unbalanced classes.

4.4. Limitations and Future Research

The sampling design is an important step to optimize the results of DSM. The inclusion of clustered sampling locations can result in biased predictions and high uncertainties [66,75]. Biases might also arise from the fact that samples of the LUCAS campaign and the soil databases are not harmonized as they were collected in different years and with different sampling protocols. SOC, as a dynamic soil property, is exposed to constant change and it is arguable whether the observed period of 1984–2014 is appropriate for accurate predictions. These factors illustrate the fact that it is necessary to optimize the models and improve the overall results. However, the main goal of this study is not to maximize the prediction accuracy of the models but to compare the transferability of stable covariate groups. Future studies should be directed to test the spatial as well as temporal transferability of remote sensing for the prediction of SOC dynamics. To guarantee comparability, harmonized soil data should be used to optimize the spatial representation and reduce biases. Recently, efforts have been made to improve the sample coverage in Germany with a nationwide Agricultural Soil Inventory (BZE-LW) to improve the quality of future DSM projects [71,76].

Unlike linear regressions, RF models and other tree-based algorithms have problems with the prediction of values outside the range of the training data [32]. It has been shown that ensemble models based on different types of learners produce better results for unknown values and should therefore be preferred for spatial or temporal extrapolation [32,61]. The next step should therefore be to test the transferability of covariates in the context of ensemble models which also showed promising results for the prediction of SOC [10]. The success of DSM extrapolation highly depends on the similarity of biophysical conditions in the donor and extended modeling area [28]. The MESS (Figure 11) and the distribution of the soil types (Table 4) illustrate that there are significant differences in the biophysical conditions of both states which influence the prediction accuracy as well as the transferability. Taking this into account, further research should be conducted to describe the similarity of different regions and define thresholds and conditions under which models can be successfully transferred [77]. Additionally, the similarity can be increased by defining suitable conditions to reduce the model area. In the case of this study, this can be done by removing certain soil types (e.g., Gleysols) which introduce extreme values and decrease the model-similarity of both states.

4.5. Recommendations for the Transfer and Extrapolation of SOC-Models

The results of our study show that there are significant differences in the transferability of covariate groups. The extrapolation of DSM models can be split into two phases to increase the number of available soil samples while still accounting for the predictive power of non-transferable covariates. For the first step, only covariate groups with a high degree of transferability should be used to train the models. These include information derived from EO data like soil reflectance composites or DEM attributes. This way, soil samples from outside the research can be included to improve the model accuracy (Figure 7). Nevertheless, the additional soil samples should be selected to reflect the natural conditions and to be as similar as possible to the research area. The proportion of training data from inside and outside the research area should be chosen to maximize the positive effect of the additional training data. For this study, the biggest accuracy increase was found when around 80% of soil samples came from outside the research area. A similar

proportion was found in the study of Zhao et al. [28] who applied a reverse k-fold cross-validation to estimate how many samples are necessary to maximize the model accuracy in an extended research area. Once the highly transferable covariates were used to predict the soil attributes, further covariates can be used to correct and improve the results. This can be realized by converting the results of the previous step into a covariate layer which can subsequently be combined with other covariates like soil legacy data or climate parameters. For this step, it is important to only include soil samples from the actual research area which corresponds to 20% of the total training samples in this case. This is necessary for including the remaining covariates that are not transferable and therefore do not produce good results with additional training samples from outside the research area. This process can optimize the prediction accuracy while minimizing the number of necessary soil samples in the research area.

5. Conclusions

The transferability of different covariate groups for the prediction of SOC in two federal states in Southern Germany was tested. The results of this experiment clearly show the importance of EO data for transferring or extrapolating DSM models regarding SOC prediction. Covariates derived from a soil reflectance composite showed the highest degree of transferability. Models based on multispectral EO data and DEM derivatives were improved by soil samples collected outside the research area. This effect was more pronounced the fewer samples inside the research area were available. A similar trend was found for terrain covariates and combined models. These findings are an important step to improve the precision of DSM models in regions with low availability of soil samples. However, the results of the non-transferred models suggest that covariates from soil legacy data and climate parameters are also important for SOC prediction. Therefore, it is proposed to split the modeling process to utilize the predictive power of transferable and non-transferable covariates. First, the EO covariates can be used to transfer DSM models to areas with low availability of soil samples. Second, additional non-transferable covariates can be combined with local samples of the research area and used to correct the results. Future work should focus on the dissimilarity of donor and recipient areas and define conditions to improve the results of transferred DSM models.

Author Contributions: Conceptualization, T.B. and T.S.; methodology, T.B., T.S. and S.Z.; software, T.B.; validation, T.B., S.S., R.T.-M. and T.S.; formal analysis, T.B., S.S., R.T.-M. and T.S.; investigation, T.B.; resources, M.B. and S.Z.; data curation, T.B., M.B. and S.Z.; writing—original draft preparation, T.B.; writing—review and editing, M.B., S.S., R.T.-M., S.Z. and T.S.; visualization, T.B.; supervision, T.S. and S.S.; project administration, T.B., T.S. and S.S.; funding acquisition, T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was founded by the German Research Foundation (DFG) through the Collaborative Research Center (SFB 1070) "Resource Cultures" (subprojects Z, S and B02), the DFG Cluster of Excellence "Machine Learning—New Perspectives for Science", EXC 2064/1, project number 390727645, and the DFG project "MLTRANS—Transferability of machine learning models for digital soil mapping", project number SCHO 739/21-1. We acknowledge the support from the Open Access Publishing Fund of the University of Tübingen.

Data Availability Statement: The data are not publicly available due to privacy restrictions.

Acknowledgments: We thank the Bavarian Environment Agency (LfU) for providing access to the soil database. We also thank the Japan Aerospace Exploration Agency (JAXA) for providing the Digital Elevation Model.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lal, R. Soil health and carbon management. *Food Energy Secur.* **2016**, *5*, 212–222. [CrossRef]
2. Lal, R. Global potential of soil carbon sequestration to mitigate the greenhouse effect. *Crit. Rev. Plant Sci.* **2003**, *22*, 151–184. [CrossRef]

3. Janzen, H.H. The soil carbon dilemma: Shall we hoard it or use it? *Soil Biol. Biochem.* **2006**, *38*, 419–424. [CrossRef]
4. Smith, P.; Falloon, P.; Kutsch, W.L. The role of soils in the Kyoto Protocol. In *Soil Carbon Dynamics: An Integrated Methodology*; Cambridge University Press: Cambridge, UK, 2010; pp. 245–256. ISBN 9780511711794.
5. Van Wesemael, B.; Chabrillat, S.; Wilken, F. High-spectral resolution remote sensing of soil organic carbon dynamics. *Remote Sens.* **2021**, *13*, 1293. [CrossRef]
6. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
7. Taghizadeh-Mehrjardi, R.; Khademi, H.; Khayamim, F.; Zeraatpisheh, M.; Heung, B.; Scholten, T. A Comparison of Model Averaging Techniques to Predict the Spatial Distribution of Soil Properties. *Remote Sens.* **2022**, *14*, 472. [CrossRef]
8. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.; Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, *2018*, e5518. [CrossRef]
9. Fathizad, H.; Taghizadeh-Mehrjardi, R.; Hakimzadeh Ardakani, M.A.; Zeraatpisheh, M.; Heung, B.; Scholten, T. Spatiotemporal Assessment of Soil Organic Carbon Change Using Machine-Learning in Arid Regions. *Agronomy* **2022**, *12*, 628. [CrossRef]
10. Taghizadeh-Mehrjardi, R.; Schmidt, K.; Amirian-Chakan, A.; Rentschler, T.; Zeraatpisheh, M.; Sarmadian, F.; Valavi, R.; Davatgar, N.; Behrens, T.; Scholten, T. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sens.* **2020**, *12*, 1095. [CrossRef]
11. Behrens, T.; Scholten, T. Chapter 25 A Comparison of Data-Mining Techniques in Predictive Soil Mapping. In *Digital Soil Mapp.ing—An Introductory Perspective*; Elsevier: Amsterdam, The Netherlands, 2006; pp. 353–617. ISBN 9780444529589.
12. Rentschler, T.; Bartelheim, M.; Behrens, T.; Díaz-Zorita Bonilla, M.; Teuber, S.; Scholten, T.; Schmidt, K. Contextual spatial modelling in the horizontal and vertical domains. *Sci. Rep.* **2022**, *12*, 9496. [CrossRef]
13. Emadi, M.; Taghizadeh-Mehrjardi, R.; Cherati, A.; Danesh, M.; Mosavi, A.; Scholten, T. Predicting and Mapp.ing of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sens.* **2020**, *12*, 2234. [CrossRef]
14. Rogge, D.; Bauer, A.; Zeidler, J.; Mueller, A.; Esch, T.; Heiden, U. Building an exposed soil composite processor (SCMaP) for mapp.ing spatial and temporal characteristics of soils with Landsat imagery (1984–2014). *Remote Sens. Environ.* **2018**, *205*, 1–17. [CrossRef]
15. Diek, S.; Fornallaz, F.; Schaepman, M.E.; de Jong, R. Barest Pixel Composite for agricultural areas using landsat time series. *Remote Sens.* **2017**, *9*, 1245. [CrossRef]
16. Demattê, J.A.; Safanelli, J.L.; Poppiel, R.R.; Rizzo, R.; Silvero, N.E.Q.; de Sousa Mendes, W.; Bonfatti, B.R.; Dotto, A.C.; Salazar, D.F.U.; de Oliveira Mello, F.A.; et al. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. *Sci. Rep.* **2020**, *10*, 4461. [CrossRef]
17. Safanelli, J.L.; Chabrillat, S.; Ben-Dor, E.; Demattê, J.A. Multispectral models from bare soil composites for mapp.ing topsoil properties over Europe. *Remote Sens.* **2020**, *12*, 1369. [CrossRef]
18. Vaudour, E.; Gomez, C.; Lagacherie, P.; Loiseau, T.; Baghdadi, N.; Urbina-Salazar, D.; Loubet, B.; Arrouays, D. Temporal mosaicking app.roaches of Sentinel-2 images for extending topsoil organic carbon content mapp.ing in croplands. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *96*, 102277. [CrossRef]
19. Zepp, S.; Heiden, U.; Bachmann, M.; Wiesmeier, M.; Steininger, M.; van Wesemael, B. Estimation of soil organic carbon contents in croplands of bavaria from scmap soil reflectance composites. *Remote Sens.* **2021**, *13*, 3141. [CrossRef]
20. Möller, M.; Zepp, S.; Wiesmeier, M.; Gerighausen, H.; Heiden, U. Scale-Specific Prediction of Topsoil Organic Carbon Contents Using Terrain Attributes and SCMaP Soil Reflectance Composites. *Remote Sens.* **2022**, *14*, 2295. [CrossRef]
21. Maleki, S.; Khormali, F.; Mohammadi, J.; Bogaert, P.; Bagheri Bodaghabadi, M. Effect of the accuracy of topographic data on improving digital soil mapp.ing predictions with limited soil data: An app.lication to the Iranian loess plateau. *CATENA* **2020**, *195*, 10481. [CrossRef]
22. Stumpf, F.; Schmidt, K.; Behrens, T.; Schönbrodt-Stitt, S.; Buzzo, G.; Dumperth, C.; Wadoux, A.; Xiang, W.; Scholten, T. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *J. Plant Nutr. Soil Sci.* **2016**, *179*, 499–509. [CrossRef]
23. Gehl, R.J.; Rice, C.W. Emerging technologies for in situ measurement of soil carbon. *Clim. Chang.* **2007**, *80*, 43–54. [CrossRef]
24. McKay, J.; Grunwald, S.; Shi, X.; Long, R.F. Evaluation of the Transferability of a Knowledge-Based Soil-Landscape Model. In *Digital Soil Mapping*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 165–178.
25. Malone, B.P.; Jha, S.K.; Minasny, B.; McBratney, A.B. Comparing regression-based digital soil mapp.ing and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma* **2016**, *262*, 243–253. [CrossRef]
26. De Arruda, G.P.; Demattê, J.; Chagas, C.; Fiorio, P.R.; e Souza, A.B.; Fongaro, C.T. Digital soil mapp.ing using reference area and artificial neural networks. *Sci. Agric.* **2016**, *73*, 266–273. [CrossRef]
27. Wolski, M.S.; Dalmolin, R.S.D.; Flores, C.A.; Moura-Bueno, J.M.; ten Caten, A.; Kaiser, D.R. Digital soil mapp.ing and its implications in the extrapolation of soil-landscape relationships in detailed scale. *Pesqui. Agropecu. Bras.* **2017**, *52*, 633–642. [CrossRef]
28. Zhao, Z.; Yang, Q.; Sun, D.; Ding, X.; Meng, F.R. Extended model prediction of high-resolution soil organic matter over a large area using limited number of field samples. *Comput. Electron. Agric.* **2020**, *169*, 105172. [CrossRef]
29. Neyestani, M.; Sarmadian, F.; Jafari, A.; Keshavarzi, A.; Sharififar, A. Digital mapp.ing of soil classes using spatial extrapolation with imbalanced data. *Geoderma Reg.* **2021**, *26*, e00422. [CrossRef]

30. Machado, D.F.T.; Silva, S.H.G.; Curi, N.; de Menezes, M.D. Soil type spatial prediction from random forest: Different training datasets, transferability, accuracy and uncertainty assessment. *Sci. Agric.* **2019**, *76*, 243–254. [CrossRef]
31. Taghizadeh-Mehrjardi, R.; Sheikhpour, R.; Zeraatpisheh, M.; Amirian-Chakan, A.; Toomanian, N.; Kerry, R.; Scholten, T. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* **2022**, *426*, 116094. [CrossRef]
32. Hengl, T. Extrapolation is Tough for Trees (Tree-Based Learners), Combining Learners of Different Type Makes It Less Tough. 2021. Available online: https://medium.com/nerd-for-tech/extrapolation-is-tough-for-trees-tree-based-learners-combining-learners-of-different-type-makes-659187a6f58d (accessed on 19 December 2022).
33. Zeraatpisheh, M.; Garosi, Y.; Reza Owliaie, H.; Ayoubi, S.; Taghizadeh-Mehrjardi, R.; Scholten, T.; Xu, M. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *CATENA* **2022**, *208*, 105723. [CrossRef]
34. BGR. Geologische Übersichtskarte der Bundesrepublik Deutschland 1: 200 000 (GÜK200). 2021. Available online: https://www.bgr.bund.de/DE/Themen/Sammlungen-Grundlagen/GG_geol_Info/Karten/Deutschland/GUEK200/guek200_inhalt.html (accessed on 19 December 2022).
35. BGR. Bodenübersichtskarte 1:200.000 (BÜK200). 2018. Available online: https://www.bgr.bund.de/DE/Themen/Boden/Informationsgrundlagen/Bodenkundliche_Karten_Datenbanken/BUEK200/buek200_node.html (accessed on 19 December 2022).
36. Jones, A.; Fernández-Ugalde, O.; Scarpa, S. *LUCAS 2015 Topsoil Survey: Presentation of Dataset and Results*; Publications Office of the European Union: Luxembourg, 2020; ISBN 9789276210801.
37. GDAL/OGR Contributors. {GDAL/OGR} Geospatial Data Abstraction Software Library. 2022. Available online: https://gdal.org (accessed on 19 December 2022).
38. Behrens, T.; Schmidt, K.; Viscarra Rossel, R.A.; Gries, P.; Scholten, T.; MacMillan, R.A. Spatial modelling with Euclidean distance fields and machine learning. *Eur. J. Soil Sci.* **2018**, *69*, 757–770. [CrossRef]
39. Møller, A.B.; Beucher, A.M.; Pouladi, N.; Greve, M.H. Oblique geographic coordinates as covariates for digital soil mapping. *SOIL* **2020**, *6*, 269–289. [CrossRef]
40. Zepp, S.; Jilge, M.; Metz-Marconcini, A.; Heiden, U. The influence of vegetation index thresholding on EO-based assessments of exposed soil masks in Germany between 1984 and 2019. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 366–381. [CrossRef]
41. JAXA EORC. ALOS Global Digital Surface Model (DSM). *ALOS World*, 2021; 1.2, 1–21.
42. Conrad, O.; Bechtel, B.; Bock, M.; Dietrich, H.; Fischer, E.; Gerlitz, L.; Wehberg, J.; Wichmann, V.; Böhner, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **2015**, *8*, 1991–2007. [CrossRef]
43. Hengl, T.; MacMillan, R.A. *Predictive Soil Mapping with R*; OpenGeoHub Foundation: Wageningen, The Netherlands, 2019; pp. 1–370. ISBN 978-0-359-30635-0.
44. Behrens, T.; Schmidt, K.; MacMillan, R.A.; Viscarra Rossel, R.A. Multiscale contextual spatial modelling with the Gaussian scale space. *Geoderma* **2018**, *310*, 128–137. [CrossRef]
45. Deutscher Wetterdienst. Open-Data-Server des Deutschen Wetterdienstes (DWD). 2022. Available online: https://opendata.dwd.de/ (accessed on 19 December 2022).
46. Hengl, T. GSIF: Global Soil Information Facilities. 2020. Available online: https://CRAN.R-project.org/package=GSIF (accessed on 19 December 2022).
47. BGR. Gruppen der Bodenausgangsgesteine 1:5.000.000 (BAG5000). 2007. Available online: https://www.bgr.bund.de/DE/Themen/Boden/Informationsgrundlagen/Bodenkundliche_Karten_Datenbanken/Themenkarten/BAG5000/bag5000_node.html (accessed on 19 December 2022).
48. BGR. Geomorphographische Einheiten der Bundesrepublik Deutschland. 2001. Available online: https://www.bgr.bund.de/DE/Themen/Boden/Bilder/Bod_Neuausrichtung_Bodenkartierung_abb1_g_720.html (accessed on 19 December 2022).
49. Ad-hoc-Arbeitsgruppe Boden. *Bodenkundliche Kartieranleitung KA5*; Schweizerbart Science Publishers: Stuttgart, Germany, 2006; ISBN 9783510959204.
50. Elith, J.; Phillips, S.J.; Hastie, T.; Dudík, M.; Chee, Y.E.; Yates, C.J. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **2011**, *17*, 43–57. [CrossRef]
51. Meyer, H.; Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [CrossRef]
52. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
53. Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
54. LeDell, E.; Gill, N.; Aiello, S.; Fu, A.; Candel, A.; Click, C.; Kraljevic, T.; Nykodym, T.; Aboyoun, P.; Kurka, M.; et al. h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. 2021. Available online: https://CRAN.R-project.org/package=h2o (accessed on 19 December 2022).
55. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. 2020. Available online: https://cran.r-project.org/package=ggpubr (accessed on 19 December 2022).
56. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Soft.* **2017**, *77*, 1–17. [CrossRef]
57. Vaudour, E.; Gholizadeh, A.; Castaldi, F.; Saberioon, M.; Borůvka, L.; Urbina-Salazar, D.; Fouad, Y.; Arrouays, D.; Richer-de-Forges, A.C.; Biney, J.; et al. Satellite Imagery to Map Topsoil Organic Carbon Content over Cultivated Areas: An Overview. *Remote Sens.* **2022**, *14*, 2917. [CrossRef]

58. Dvorakova, K.; Heiden, U.; van Wesemael, B. Sentinel-2 Exposed Soil Composite for Soil Organic Carbon Prediction. *Remote Sens.* **2021**, *13*, 1791. [CrossRef]
59. Urbina-Salazar, D.; Vaudour, E.; Baghdadi, N.; Ceschia, E.; Richer-de-Forges, A.C.; Lehmann, S.; Arrouays, D. Using Sentinel-2 Images for Soil Organic Carbon Content Mapping in Croplands of Southwestern France. The Usefulness of Sentinel-1/2 Derived Moisture Maps and Mismatches between Sentinel Images and Sampling Dates. *Remote Sens.* **2021**, *13*, 5115. [CrossRef]
60. Žížala, D.; Minařík, R.; Zádorová, T. Soil Organic Carbon Mapp.ing Using Multispectral Remote Sensing Data: Prediction Ability of Data with Different Spatial and Spectral Resolutions. *Remote Sens.* **2019**, *11*, 2947. [CrossRef]
61. Bonannella, C.; Hengl, T.; Heisig, J.; Parente, L.; Wright, M.N.; Herold, M.; de Bruin, S. Forest tree species distribution for Europe 2000–2020: Mapping potential and realized distributions using spatiotemporal machine learning. *PeerJ* **2022**, *10*, e13728. [CrossRef] [PubMed]
62. Fathololoumi, S.; Vaezi, A.R.; Alavipanah, S.K.; Ghorbani, A.; Saurette, D.; Biswas, A. Improved digital soil mapp.ing with multitemporal remotely sensed satellite data fusion: A case study in Iran. *Sci. Total Environ.* **2020**, *721*, 137703. [CrossRef] [PubMed]
63. Minasny, B.; McBratney, A. Why calculating RPD is redundant. *Pedometron* **2013**, *33*, 14–15.
64. Wiesmeier, M.; Hübner, R.; Barthold, F.; Spörlein, P.; Geuß, U.; Hangen, E.; Reischl, A.; Schilling, B.; von Lützow, M.; Kögel-Knabner, I. Amount, distribution and driving factors of soil organic carbon and nitrogen in cropland and grassland soils of southeast Germany (Bavaria). *Agric. Ecosyst. Environ.* **2013**, *176*, 39–52. [CrossRef]
65. Ma, T.; Brus, D.J.; Zhu, A.-X.; Zhang, L.; Scholten, T. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. *Geoderma* **2020**, *370*, 114366. [CrossRef]
66. Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Model. Softw.* **2018**, *101*, 1–9. [CrossRef]
67. Castaldi, F.; Chabrillat, S.; Don, A.; van Wesemael, B. Soil organic carbon mapp.ing using LUCAS topsoil database and Sentinel-2 data: An app.roach to reduce soil moisture and crop residue effects. *Remote Sens.* **2019**, *11*, 2121. [CrossRef]
68. Dvorakova, K.; Shi, P.; Limbourg, Q.; van Wesemael, B. Soil Organic Carbon Mapping from Remote Sensing: The Effect of Crop Residues. *Remote Sens.* **2020**, *12*, 1913. [CrossRef]
69. Guo, L.; Sun, X.; Fu, P.; Shi, T.; Dang, L.; Chen, Y.; Linderman, M.; Zhang, G.; Zhang, Y.; Jiang, Q.; et al. Mapp.ing soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma* **2021**, *398*, 115118. [CrossRef]
70. Guanter, L.; Kaufmann, H.; Foerster, S.; Brosinsky, A.; Wulf, H.; Bochow, M.; Boesche, N.; Brell, M.; Buddenbaum, H.; Chabrillat, S.; et al. *Environmental Mapping and Analysis Program (EnMAP) EnMAP Science Plan 2016*; EnMAP Consortium: Potsdam, Germany, 2016. [CrossRef]
71. Jacobs, A.; Flessa, H.; Don, A.; Heidkamp, A.; Prietz, R.; Dechow, R.; Gensior, A.; Poeplau, C.; Riggers, C.; Schneider, F.; et al. *Landwirtschaftlich genutzte Böden in Deutschland: Ergebnisse der Bodenzustandserhebung*; Thünen Report No. 64.; Johann Heinrich von Thünen-Institut: Braunschweig, Germany, 2018; Volume 64. [CrossRef]
72. Vohland, M.; Ludwig, B.; Seidel, M.; Hutengs, C. Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. *Geoderma* **2022**, *405*, 115426. [CrossRef]
73. Gholizadeh, A.; Neumann, C.; Chabrillat, S.; van Wesemael, B.; Castaldi, F.; Borůvka, L.; Sanderman, J.; Klement, A.; Hohmann, C. Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions. *Soil Tillage Res.* **2021**, *211*, 105017. [CrossRef]
74. Hong, Y.; Chen, S.; Chen, Y.; Linderman, M.; Mouazen, A.M.; Liu, Y.; Guo, L.; Yu, L.; Liu, Y.; Cheng, H.; et al. Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest. *Soil Tillage Res.* **2020**, *199*, 104589. [CrossRef]
75. Hengl, T.; Parente, L.; Bouasria, A.; Wheeler, I. *Spatial Sampling and Resampling for Machine Learning*; Zenodo: Geneva, Switzerland, 2022. [CrossRef]
76. Sakhaee, A.; Gebauer, A.; Ließ, M.; Don, A. Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms. *SOIL* **2022**, *8*, 587–604. [CrossRef]
77. Roudier, P.; Odgers, N.; Carrick, S.; Eger, A.; Hainsworth, S.; Beaudette, D. Soilscapes of New Zealand: Pedologic diversity as organised along environmental gradients. *Geoderma* **2022**, *409*, 115637. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.