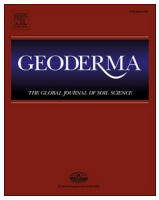
Geoderma 444 (2024) 116867

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



# Soil organic carbon mapping utilizing convolutional neural networks and Earth observation data, a case study in Bavaria state Germany

Nikolaos Tziolas [a,*], Nikolaos Tsakiridis [b], Uta Heiden [c], Bas van Wesemael [d]

[a] Southwest Florida Research and Education Center, Department of Soil, Water and Ecosystem Sciences, Institute of Food and Agricultural Sciences, University of Florida, 2685 State Rd 29N, Immokalee, FL 34142, USA
[b] Spectra Lab Group, Laboratory of Remote Sensing, Spectroscopy, and GIS, Department of Agriculture, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[c] German Aerospace Center, The Remote Sensing Technology Institute, Department of Photogrammetry and Image Analysis, Oberpfaffenhofen, D-82234 Wessling, Germany
[d] Earth and Life Institute, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Handling Editor: B. Minasny

*Keywords:*
Deep learning
Common agricultural policy
Explainable artificial intelligence
Digital soil mapping

ABSTRACT

The Copernicus Sentinel-2 multispectral imagery data may be aggregated to extract large-scale, bare soil, reflectance composites, which enable soil mapping applications. In this paper, this approach was tested in the German federal state of Bavaria, to provide estimations for soil organic carbon (SOC). Different temporal ranges were considered for the generation of the composites, including multi-annual and seasonal ranges. A novel multi-channel convolutional neural network (CNN) is proposed. By leveraging the advantages of deep learning techniques, it utilizes complementary information from different spectral pre-treatment techniques. The SOC predictions indicated little dissimilarity amongst the different composites, with the best performance attained for the six-year composite containing only spring months (RMSE = 12.03 g C · $kg^{-1}$, $R^2$ = 0.64, RPIQ = 0.89). It has been demonstrated that these outcomes outperform other well-known machine learning techniques. An ablation analysis was accordingly performed to evaluate the interplay of the CNN's different components to disentangle the advantages of each aspect of the proposed framework. Finally, a DUal inPut deep LearnIng architecture, named DUPLICITE, is proposed, which concatenates deep spectral features derived from the CNN mentioned earlier, as well as topographical and environmental covariates through an artificial neural network (ANN) to exploit their complementarity. The proposed approach was demonstrated to provide an improvement in the overall prediction performance (RMSE = 11.60 gC · $kg^{-1}$, $R^2$ = 0.67, RPIQ = 0.92).

## 1. Introduction

Monitoring of the soil ecosystem and health should be improved by adopting integrated and multi-dimensional approaches employing Earth Observation (EO) techniques, to safeguard the natural capital of soils. This becomes increasingly clearer due to the impact of persistent environmental changes taking place in a continuously volatile environment of reduced resources. Such an improvement will enable the formulation of evidence based conservation policy recommendations (Montanarella, 2015) and sustainable services, such as carbon smart agriculture (Paul et al., 2023).

Many studies demonstrated the capacity of contemporary multi-spectral spaceborne EO time series for topsoil mapping, showcasing the advantages of both the short revisit times and the unprecedented spatial resolution (≤10 m). A review by Tziolas et al. (2021), summarized the EO-based soil monitoring and mapping methodologies; therein, both the shortcomings and the advantages of current spaceborne systems were detailed. Moreover, to overcome some of the existing limitations and caveats, a brief overview of novel Artificial Intelligence (AI) techniques and information and communication technology trends was provided. Some notable examples for the estimation of soil properties include the use of multispectral imager such as Landsat 8 (Asgari et al., 2020; Gasmi et al., 2021) and Sentinel-2 (Azizi et al., 2023), as well as the PRISMA hyperspectral imager (Mzid et al., 2022).

Satellites with high revisit times and advancements in data analytics enable the construction of multi-date soil reflectance composites (SRCs) that can be used for predicting topsoil properties over croplands (Rizzo et al., 2023). The widely used Normalized Difference Vegetation Index

* Corresponding author.
*E-mail addresses:* ntziolas@ufl.edu (N. Tziolas), tsakirin@auth.gr (N. Tsakiridis), uta.heiden@dlr.de (U. Heiden), bas.vanwesemael@uclouvain.be (B. van Wesemael).

https://doi.org/10.1016/j.geoderma.2024.116867
Received 17 April 2023; Received in revised form 7 March 2024; Accepted 25 March 2024
Available online 28 March 2024

(NDVI) is employed to filter out vegetated pixels through specific threshold values, while the Normalized Burn Ratio and the Bare Soil Index (Dvorakova et al., 2020; Mzid et al., 2021) have been used to eliminate the effect of soil moisture and crop residues. Previous works explored custom indices with thresholds derived heuristically or via expert knowledge (Loiseau et al., 2019; Žížala et al., 2019). Considering these recent findings, we can conclude that further knowledge is required about the diversity of index based thresholds over large areas and their impact on the predictive performance of the AI models. In that regard, Heiden et al. (2022) recently proposed a novel histogram separation methodology to derive thresholds from Landsat imagery data making use of the soil composite mapping processor (SCMaP, Rogge et al., 2018).

SRCs have demonstrated their capacity to provide topsoil monitoring (Vaudour et al., 2021); however, the commonly employed machine learning (ML) algorithms in EO-driven cropland soil monitoring do not fully utilize the spectral information. Recently, the advent of deep learning (DL) methods has provided a significant advance in soil spectral analysis by incorporating multiple spectral sources and integrating their information through the operation of convolution (Tsakiridis et al., 2020). Compared to previous studies that used indices as additional predictors in conventional ML algorithms (Lamichhane et al., 2021; Shi et al., 2022), DL models allow streamlining the pipeline, reducing the human expertise in feature engineering and enabling the model to learn intricate patterns directly from the raw data. Nevertheless, a challenge in EO data analysis concerns the effective fusion of optical data (such as multi-spectral datasets) with environmental and topographical data from other sensors (i.e., two information sources that are complementary for digital soil mapping). Studies so far have exploited common ML approaches with a simple combination of spectral input data and environmental covariates (Zeraatpisheh et al., 2022). Combinations of heterogeneous EO data, such as optical and radar, can help to enhance the scene selection for building composites (Urbina-Salazar et al., 2021), but distinct improvements of the topsoil predictions have still to be proven.

The overarching objective of our work was to introduce an innovative framework employing a multi-channel convolutional neural network (CNN) for soil organic carbon (SOC) content prediction utilizing the SRCs from Senitnel-2. The model could employ, as input, either the mean "masked" bare soil reflectance spectra (single approach) over time, or simultaneously utilize the reflectance spectra and spectra derived from common pre-treatment methods (multi-channel approach). An analysis and comparison of each component's impact on the estimation accuracy with different temporal composites took place. A component analysis using a local adaptive error-correction approach was also carried out. Lastly, we introduced DUPLICITE, which stands for DUal inPut deep LearnIng architecture that simultaneously combined optical and environmental data (i.e., covariates) through two distinct neural network branches.

## 2. Materials and methods

The methodological approach consists of two discrete steps: i) the processing involves generating the SRCs and ii) preparing climatic and topographic variables, which are fed into the DL models to estimate SOC. An overview of the proposed workflow is illustrated in Fig. 1.

### 2.1. Test site

Our methodology was validated in the state of Bavaria, Germany (70,553 km$^2$, Fig. 2a). The test site included mainly croplands showcasing great SOC variability (Fig. 2b). It should be noted that several drained peatbog areas (e.g., the Donaumoos) may be found in the South. More information about the test site may be found in (Wiesmeier et al., 2020). Luvisols and Cambisols are the most prevalent soil classes in the study area (Working Group WRB, 2015, Fig. 2c).

### 2.2. Data

Table 1 provides an overview of the dataset used in this work.

#### 2.2.1. Optical data from Sentinel-2

To develop composite products, Sentinel-2 imagery data was used. The entire study area is covered by nine and three tiles from UTM (Universal Transverse Mercator) zones 32 and 33, respectively. (Vermote et al., 2016) indicated that a reliable approximation of atmospheric parameters, such as aerosols and water vapor, using a scene-based analysis, cannot be guaranteed. Therefore, we retained only images ranging from March to October, due to the low sun zenith angle (>70°) occurred during the winter months. Moreover, we constrained the image selection with a cloud coverage of < 80 %. The three 60 m Sentinel-2 bands (B1, B9, and B10) were excluded from the following analysis. Further, the B2-B4 and B8 were resampled to 20 m pixel size, applying the nearest neighbor method. In total, 2551 single images were used for creating SRCs collecting bare soil pixels from varying years and seasons. The MAJA atmospheric correction algorithm (Hagolle et al., 2017, 2010) was used to obtain the L2A products, along with the generation of cloud, shadow, and snow information masks used for producing the SRCs.

#### 2.2.2. Soil organic carbon data

To achieve an accurate and detailed topsoil mapping, various sources were used to identify all the available reference SOC data in the area. Thus, two soil sample databases from federal authorities, namely the Bavarian Environment Agency (LFU) and the Bavarian Research Center for Agriculture (LFL), as well as the available samples from the Land Use and Coverage Area Frame Survey (LUCAS) topsoil database were utilized (Orgiazzi et al., 2018).

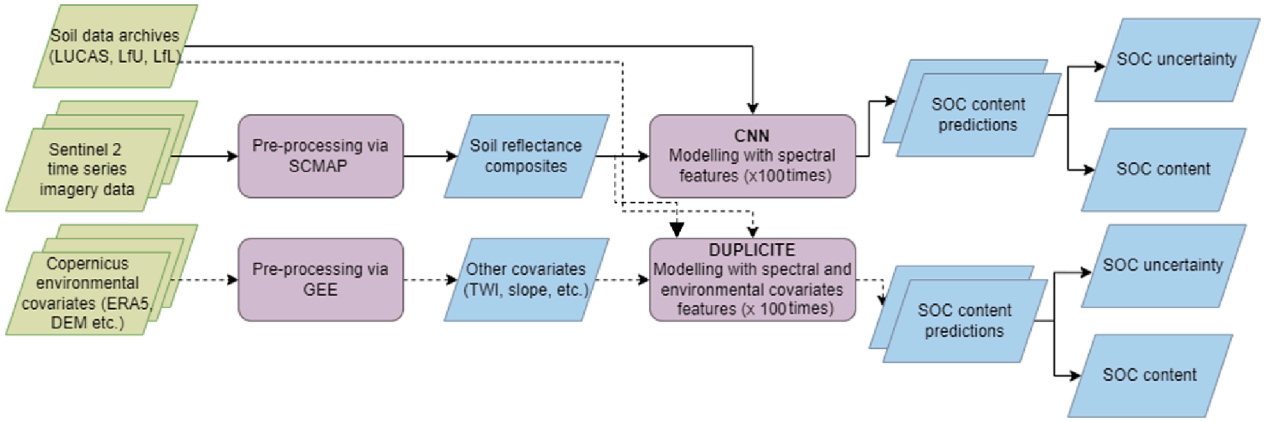



**Fig. 1.** Overview of the proposed EO based SOC mapping framework.

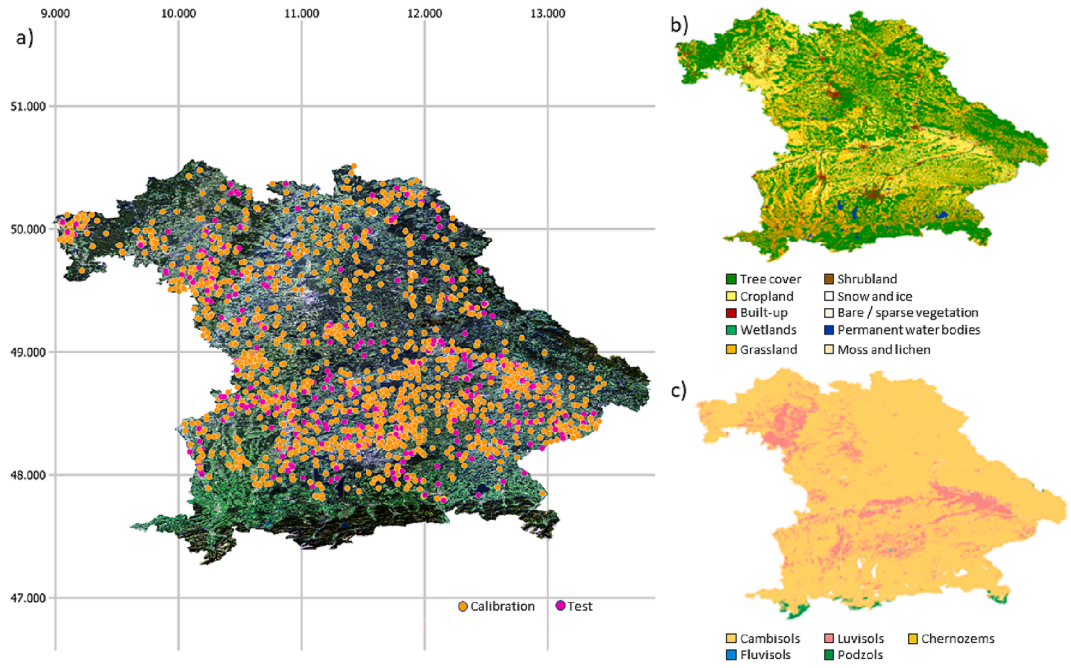



**Fig. 2.** Test site in Bavaria. a) Calibration (orange) and test (purple) sampling points displayed on a cloud free mosaic of Sentinel-2 data from 2017; b) land use product from ESA WORLDCOVER and c) soil classification map.

**Table 1**
Summary of data generated by EO sources and soil data used in this study.

| Label | Data description | Features |
|---|---|---|
| Sentinel-2 | Archive of L2A multi-spectral data | 10 Spectral bands (B2-8, B8A, B11 and B12) with spatial resolution between 10 and 20 m (all resampled to 20 m). |
| Soil data | Land Use and Cover Area Frame Survey 2015 topsoil spectral database (LUCAS, only mineral croplands), and from the Bavarian Environment Agency (LFU) and Bavarian Research Center for Agriculture (LFL) regional soil databases | Spectral Signatures (400–2500 nm) and SOC content; gravimetric content of organic carbon in the fine earth fraction of the soil (g C · $kg^{-1}$) |
| Corine land cover | Land cover spatial product | Urban (111 and 112); Arable land (211, 212 and 213); Deciduous (311); Grasslands (231, 321 and 322) classes |
| Climate data | Precipitation and temperature from the European Centre for Medium-Range Weather Forecasts (ECMWF) global climatic re-analysis (2000–2021) | 2 m air temperature (min, max and mean) and total precipitation |
| Digital Elevation Model (DEM) | DEM over European territory | Land surface elevation from the Copernicus dataset, and the derivatives (slope and topographic wetness index) |

The LFU database included 885 samples for a depth of approximately 0–35 cm. Soil material was gathered as a composite sample from eight sub-locations surrounding the main soil profile to account for the small-scale heterogeneity. As mentioned in the earlier work of Wiesmeier et al. (2012), the samples were taken between 1984 and 2014 from sites distributed equally throughout Bavaria. The LFL dataset comprised information derived from Bavaria's permanent soil observation program. A total of 401 samples were collected from various locations in the Bavaria state since 1985, using a soil auger at a depth of 0 to 10 cm. For each site, five random sampling locations were selected. Within a 1.5 m radius at each location, six soil samples were gathered. The SOC contents of all six samples from each of the five locations were averaged, resulting in a single SOC content value. In that regard, it should be noted that there was no significant change in SOC content in almost all permanent soil observation sites based on the work of Kühnel et al. (2020). In the recent study by Zepp et al. (2021), the same dataset was used, and it is mentioned that there are no significant changes in the SOC values across the years. Therefore, samples from all databases have been used for SOC model calibration. The determination of SOC contents for both LFL and LFU databases was performed by dry combustions using a CN elemental analyzer. It is worth noting that utilizing legacy data comes with a drawback, since the sampling schemes are not optimally distributed. The LUCAS soil samples were collected in 2009 from distinct points within the area of interest, following the protocol described in the work of Tóth et al. (2013), while SOC content was measured after dry combustion with an elemental analyzer following the ISO 10694:1995 protocol (ISO, 1995). Considering the three databases, more than 1500 topsoil samples were available (Table 2), with the SOC contents ranging from 2.6 to 180 g C · $kg^{-1}$.

### 2.2.3. *Topographic and climatic data*

Climatic variable datasets (precipitation and temperature) in gridded form were obtained from ERA5 through the Google Earth Engine (Gorelick et al., 2017). The ERA5 dataset combines a global climate model with *in situ* and satellite observations (Hersbach et al., 2020). Statistics,

**Table 2**
Number of soil samples and information about the year of sampling and the contributing organizations.

| Acronym | Contributor | Sampling Year | Observations |
|---|---|---|---|
| LFU | Bavarian Environment Agency | 1984 to 2014 | 885 |
| LFL | Bavarian Research Center for Agriculturen | 1985 | 401 |
| LUCAS | European Soil Data Centre — Land Use/Cover Area frame statistical survey | 2015 | 437 |

including average, minimum, maximum, and cumulative values were computed for the period 2000–2021. For precipitation we considered only the cumulative values. The digital elevation model (DEM) from the European Union's Copernicus land monitoring service was employed, projected to the WGS84 coordinate reference system, and then imported into GRASS GIS to compute the topographical wetness index (Beven and Kirkby, 1979), and extract the elevation and slope values.

#### 2.2.4. Corine land cover

The Corine land cover used in our study since it represents an inventory of land use and land cover classes at a pan European level. The Corine Land Cover database was utilized for threshold-based compositing approach (see Section 2.3), to enable the separation of bare soils from other spectrally similar land cover types without using other auxiliary information. In this context, we focused on targeted agricultural related classes or classes having similar spectral characteristics across the different phenological stages. The CLC layers from 2012 and 2018 were utilized to calibrate the thresholds on temporally stable areas representative of key land covers classes. The following classes were used to mask out the region of interest, as indicated in Table 3.

Lastly, the original vector product was rasterized to match the spatial resolution of the Sentinel-2 imagery data time series.

### 2.3. Generation of soil reflectance composites

Three discrete steps are required to generate the bare soil composites: i) retrieving and storing L2A data from Sentinel-2 sensor; ii) EO data pre-processing (including atmospheric correction) as well as data cleansing (i.e., removal of duplicates); and iii) executing SCMaP. Rogge et al. (2018) provided a thorough detail of the overall pipeline and the logic behind SCMaP. The SCMaP was applied for topsoil mapping at various geographical scales (Möller et al., 2022). Herein we described it briefly, emphasizing on the modifications made to operate on Sentinel-2 data (Heiden et al., 2022).

The SCMaP has as input the multi-temporal EO data and, through a pixel-based approach, generates the bare SRC to be used in the subsequent modelling phase. We focus on the generic methodology HIstogram SEparation Threshold (HISET) for deriving spectral index thresholds to create the SRCs. In brief, HISET has been utilized to derive spectral index thresholds, that in turn are utilized to discriminate bare soils from all other land cover classes (e.g., urban areas). HISET is independent from specific indices, however, the commonly used NDVI and NBR2 indices, as well as the new index PV + IR2 (Equation (1) which combines the information from the Visible to Near Infrared (VNIR) and the Short Wave Infrared (SWIR) wavelength regions.

$$PV + IR2 = \frac{(B8 - B4)}{(B8 + B4)} + \frac{(B8 - B12)}{(B8 + B12)} \quad (1)$$

The barest maximum and minimum index composite are used to derive the $t_{max}$ and $t_{min}$ thresholds, respectively. $T_{min}$ corresponds to separating non-photosynthetic active vegetation from bare soils, while $t_{max}$ is employed to distinguish permanently non-vegetated lands (e.g., urban areas) from vegetated soils. The soil mask is generated via the intersection of both results and contains surfaces that alternate between vegetated and bare soils in the temporal dimension, which is the characteristic of cropland soils. In the final step, the $t_{min}$ index is utilized to filter out the bare soil pixels within the soil mask area. As indicated in Table 4, we tested different sets of months to remove limitations induced by the several disturbing factors on the reflectance data that exhibit reduced seasonal series marked by the driest soils on an annual basis.

**Table 3**
Corine Land Cover Classes and Corresponding Codes.

| Corine land cover class | Corine land cover codes | Name |
|---|---|---|
| Agricultural | 211, 212, 213 | crops |
| Grassland | 231, 321, 322 | grassland |
| Deciduous forest | 311 | deciduous |
| Urban | 111, 121 | urban |

**Table 4**
The settings of the different composites examined per each time range experiment.

| SRC name | Years | Months |
|---|---|---|
| SRC 2018–2020 full | 2018–2020 | March to October |
| SRC 2018–2020 spring_autumn | 2018–2020 | March to May and August to October |
| SRC 2018–2020 spring | 2018–2020 | March–May |
| SRC 2015–2020 spring_autumn | 2015–2020 | March to May and August to October |
| SRC 2015–2020 spring | 2015–2020 | March to May |

### 2.4. Deep learning spectral modelling architectures

#### 2.4.1. The proposed CNN approach

Here, we introduce the proposed multi-input CNN, which utilizes only the multispectral data as input to estimate SOC. The concept behind the approach is to leverage multiple pre-treatment techniques (i.e., pseudo absorbance conversion and the standard normal variate, SNV) simultaneously which contain complementary information. The complementarity is based on the fact that pre-treatment techniques affect the spectral signal differently (e.g., scatter correction or highlighting absorption peaks), each developing their unique input space.

The CNN architecture is composed of five different building blocks: i) the input layer, where the three spectral signatures (i.e., the initial reflectance and the two pre-processed spectra) are fed into the network as different input channels; ii) the one-dimensional convolutional layers, extracting information from the spectra through convolution; iii) the pooling layer, which down-samples the signal; iv) the flattening layer, assembling the extracted deep features into a flattened representation; and v) the fully-connected layers (also termed dense layers). The Adam optimizer (Kingma and Ba, 2015) was used to train the network with a batch size of 10. Further, the maximum learning epochs are 400, while an early stop at 40 epochs is used. Fig. 3 illustrates the overall network architecture.

The zero-padded multi-channel input is first convolved using a one-dimension convolutional layer, comprised of 16 filters using a window of length three: thus, more precisely, using a kernel of size 3 (number of input channels) × 3 (convolution window across the spectral dimension), ensuring the combination of multiple inputs. A batch normalization was applied, followed by a Leaky Rectified Linear Unit (ReLU; $\alpha = 0.01$). The ReLU also works as an activation function. It should be noted that the batch normalization normalizes the weights of a layer by shifting and scaling the activations, smoothing the optimization landscape. This leads to faster and more stable training. The filtered signals are then down-sampled by halving their dimensionality using a max pooling layer. Afterwards, a second one-dimension convolutional layer (32 filters) is employed having the same hyper-parameters as the 1st convolutional layer. The convolution results are accordingly flattened whereby a dense network produces the predictions. The dense layer makes use of two internal layers (with 32 and 8 neurons, respectively) and a Leaky ReLU activation ($\alpha = 0.01$) function. Further a L2 kernel regularizer ($\lambda_2 = 4E_{-4}$) penalizes the weights to reduce over-fitting, and one output layer of size 1 (equal to the size of the output; soil target property) making use of the tanh activation function, since it performed better than sigmoid or linear functions. Table 5 summarizes all the layers.

To show the efficacy of our approach, we compared the original version of the CNN with an architecture considering the reflectance spectra values extracted from the SRCs as a single input channel, denoted by single-CNN. The same network architecture presented in Table 5 and the optimal hyper-parameters used also in this model.

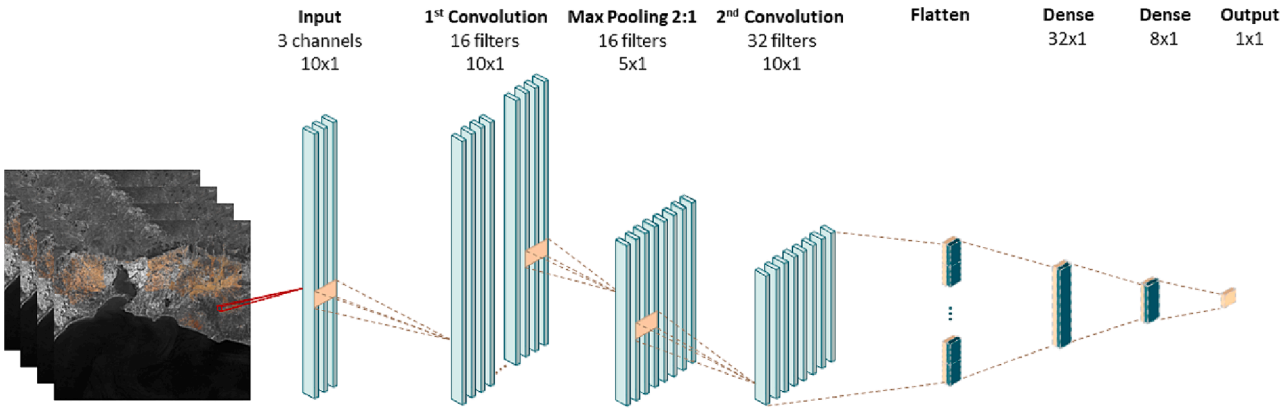



**Fig. 3.** The CNN architecture for SOC content prediction using Sentinel-2 imagery data. The orange vertical rectangles represent the dropout layers, while horizontal correspond to kernel filters.

**Table 5**
Description of the CNN network architecture along with the various layers.

| Layer type | Filters | Kernel size | Width | Activation |
|---|---|---|---|---|
| Convolutional and batch normalization | 16 | 3x1 | 160 | Leaky ReLU |
| Max pooling | – | 1x2 | 64 | – |
| Convolutional and batch normalization | 32 | 3x1 | | Leaky ReLU |
| Flatten | – | – | 10,272 | – |
| Fully connected | – | – | 32 | Leaky ReLU |
| Fully connected | – | – | 8 | Leaky ReLU |
| Fully connected | – | – | 1 | tanh |

Moreover, its results were compared with other AI algorithms (partial least squares (PLS, Wold et al., 2001); random forest (RF, Breiman, 2001); support vector machines (SVM, Drucker et al., 1997); and XGBooST (Chen and Guestrin, 2016)).

### 2.4.2. *Investigating the interpretability of the DL models*

To examine the CNN model's interpretability, we focused on the absorbance values [log10 (1/Reflectance)] that are used in the multi-channel CNN model as one of the input spectra channels. We then visualized the activations for the top and bottom 5 % patterns in each filter of the second convolutional layer, which represents the deep features automatically constructed by the network. The goal was to emphasize that the interpretability of the learned features should be the cornerstone of the CNN architectures' conceptualization and design, to shed a light into the mechanisms and inference of the model, in addition to its discrimination rules.

### 2.4.3. *Localized learning using an adaptive error correction layer*

The component presented here is inspired by the previous work of Tsakiridis et al. (2020). They proposed a local error correction mechanism in a hyperspectral dataset. A modified version is used in this paper.

The main concept of this step is to improve the predictions of the CNN. The proposed error-correction mechanism uses the $k$ closest neighbors to adapt the model's estimation and is thus considered local. As distance function to select the neighbors, the Euclidean distance in the principal component space is employed. As shown in Fig. 4, for each unknown sample the estimated SOC content $\hat{y}$ is offset through the prediction errors of its k-closest neighbors in the calibration set, thus producing the final estimation $\hat{z}$.

In brief, for the unknown patterns, the predictions by the multi-channel CNN are adjusted based on the neighbors and their known predictions errors. Utilizing the technique proposed by Tetko and Tanchuk (2002), we calculated the new estimation for every soil attribute based on the Equation (2).

$$\hat{z} = \hat{y} + \frac{\sum_{i=1}^{k} e_i \bullet F(d_i)}{\sum_{i=1}^{k} F(d_i)} \tag{2}$$

where $\hat{y}$ is the prediction, $e_i$ represents the prediction error of the calibration pattern $i$, and the $F(d_i)$ is a scaling function that utilized as input the distance between the calibration pattern $i$ and the unknown multispectral-spectral signature from Sentinel-2. To estimate the $F(d_i)$ the function proposed by Bruneau and McElroy (2006) has been utilized (Equation (3):

$$F(d_i) = \begin{cases} 1 & if\, d_i \le pl \\ exp(-(d_i - pl)^2/(2\sigma^2)) & otherwise \end{cases} \tag{3}$$

where $pl$ is a plateau below which the function takes on a value of one, while σ regulates the exponential decay. It should be noted that the goal

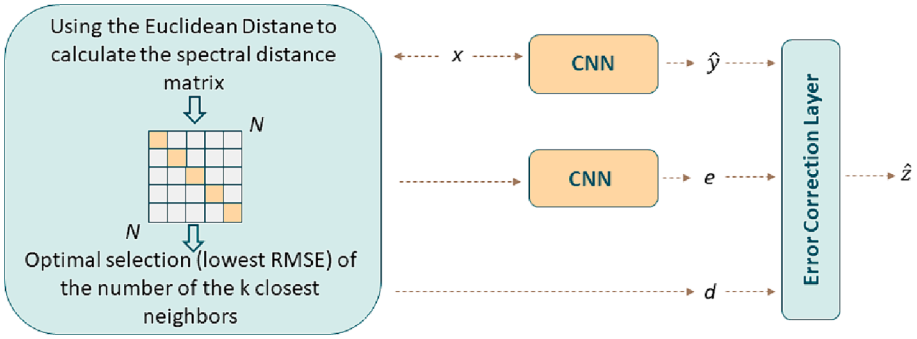



**Fig. 4.** Overview of the local error correction mechanism.

of the local adaption correction layer is to make slight adjustments to predictions, hence we opted for a single layer at the initial architecture encoding a pair of the plateau (*pl*) aforementioned and the factor σ hyperparameters for each property. This allows us to independently fine tune these values for each testing pattern, while the number of selected neighbors requires a cross validation towards the optimization of the final prediction error.

#### 2.4.4. DUPLICITE: The proposed dual input deep learning architecture

This section details the novel DUPLICITE framework, aiming to estimate soil properties using as input a combination of both the spectral values and a set of environmental and topographical covariates commonly employed in digital soil mapping. The key is that the diverse input data are ingested to two structurally divergent streams. In this context, each one is composed of a branch defining two discrete neural networks: i) a CNN and ii) an artificial neural network (ANN). Considering that the branches explore diverse information from different points of view, they supply complementary information for the regression analysis. The CNN network uses as predictors the Sentinel-2 bands and constructs the deep features as mentioned previously, whereas the ANN utilizes the environmental covariates. In that regard, each branch produces a vector of characteristics that summarizes the knowledge that was extracted. Finally, the results of both networks are concatenated in a single feature vector of descriptor used as input into a dense layer that produces the final prediction of the target variable. The proposed DUPLICITE architecture is presented in Fig. 5.

All experiments were executed on a workstation powered by an Intel® Core™ i7 11800H processor (base Frequency of 2.30 GHz and 16 GB RAM) with a CUDA-enabled NVIDIA-GeForce RTX 3060 graphics process unit. The Python Tensorflow library has been used to implement the DL methods. The average training time is around five minutes, while the production of each map requires around 30 min.

### 2.5. Model evaluation metrics

The results have been assessed using three commonly employed statistical indicators, namely the Root Mean Square Error (RMSE, Equation (4); the coefficient of determination ($R^2$, Equation (5); and the Ratio of Performance to Interquartile Range (RPIQ, Equation (6):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{i=N}(y_i - \hat{y}_i)^2}{N}} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{i=N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{i=N}(y_i - \bar{y})^2} \tag{5}$$

$$\mathrm{RPIQ} = \frac{IQ}{\mathrm{RMSE}} \tag{6}$$

where $y_i$ = observed value and $\hat{y}_i$ = predicted value, $\bar{y}$ = mean of the observed values, N = number of observations with $i = 1,2,\ldots,$ N, and *IQ* is the interquartile range of the observed values ($IQ = Q_3 - Q_1$).

Furthermore, to visualize the uncertainty of the predictions as a spatial product, we calculated (Equation (7) the 90th prediction interval (PI90) indicator as previously used by Poggio et al. (2021):

$$PI90 = q_{0.95} - q_{0.05} \tag{7}$$

The $q_{0.95}$ is the 95th and $q_{0.05}$ the 5th percentile.

### 2.6. Dataset split

The Conditioned Latin Hypercube algorithm (Minasny and McBratney, 2006) has been applied to divide the initial dataset into calibration and test sets. To guarantee that the independent test pixels are consistently chosen across all SRCs and that the results (e.g., evaluation metrics) are comparable, we chose the one with the fewest bare soil pixels from the composites presented in Table 4. Therefore, the "SRC 2018–2020 spring" dataset was split into 70–30 % to result 735 calibrations and 315 test points. In all modeling scenarios, a common test dataset of these 315 points was used. Then, the calibration dataset was divided into five folds (internal cross-validation). This method was utilized to estimate the best hyperparameters (PLS: Latent Variables = 10; RF: the minimum number of samples for terminal nodes = 10, number of

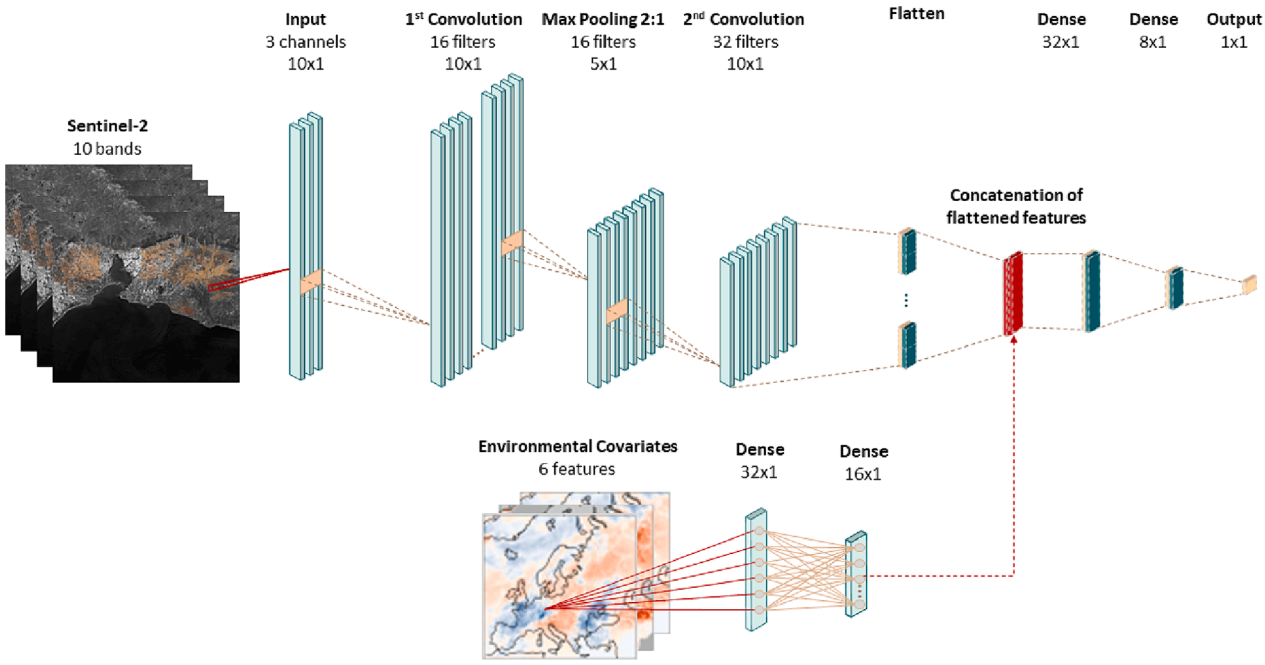



**Fig. 5.** The proposed DUPLICITE architecture for SOC prediction using optical spaceborne data (Sentinel-2) and environmental and topographical covariates; utilizing two structurally different neural networks (CNN and ANN). The red rectangle indicates the concatenation step, the orange vertical rectangles refer to dropout layers, while the horizontal ones refer to the kernel filters.

trees = 1000 and max number of levels in each decision tree = 24 SVM: Sigma = 0.01 and C = 2; XGBooST: number of rounds to run the training = 1000, step size shrinkage used in updates to prevent overfitting = 0.025, minimum loss reduction = 0 and maximum depth of a tree = 2).

## 3. Results

### 3.1. Approaches for generation of soil reflectance composites

It is worth noting that bare soil pixels mostly occur in the spring and autumn season. In brief, we can define the period ranging from March to May as the spring season, while autumn ranges from August to October. The rationale behind this selection derived from the Bavarian crops and their phenological stages and the corresponding management practices (e.g., tillage, seeding and harvesting). Moreover, it should be mentioned that the number of valid pixels also depends on the cloud coverage of the respective months.

Thus, the temporal transition between non-vegetated (e.g., urban) and vegetated pixels allowed the detection of the bare soil pixels, setting the vegetation index thresholds at 0.32 for $t_{min}$ and 1.42 for $t_{max}$ (Fig. 6). The thresholds for the five composites (Table 4) have been derived based on the available input data from 2018 − 2020 (March to October).

The performance of each SRC was evaluated by calculating the percentage of bare soils pixels within the test site and the total points with observed SOC contents contained at each SRC. We have also compared the resampled Sentinel-2 reflectance values from the LUCAS library with the SRC values of each point, using the spectral angle (Table 6). The results indicated that the difference between the "SRC 2018–2020 full" composite and the "SRC 2018–2020 spring_autumn" composite is marginal for all statistical measures (e.g., bare soils %, etc.). However, a significant difference can be noticed between the "SRC 2018–2020 full" and the "SRC 2018–2020 spring" composites. The restriction to spring months resulted in a reduction of around 40 % of the covered LUCAS points (137 instead of 237). Additionally, we estimated the spectral similarity between the SRC spectra and LUCAS spectra using the spectral angle. The results range from 0.084 for the "SRC 2018–2020 full" composite to 0.069 for the "SRC 2018–2020 spring" composite. The results indicated that although less pixels were selected by the *spring* composites, the spectral match to LUCAS improved. Considering the spring-autumn composite that was generated for a period of six years (SRC 2015–2020 spring_autumn), the performance measures were comparable with the three years composites, 32.20 % bare soil pixels and 0.082 rad spectral match.

### 3.2. Modelling approaches

The sampling points which were characterized as bare soil pixels at each SRC were used in the subsequent steps of our data analysis. As noted, we used the proposed CNN to five different SRCs to emphasize the impact of composites generated at various periods, which led to changes in topsoil conditions. Table 7 displays the results of the accuracy metrics from the various experiments. The best performance was achieved for the composite that stacks the spring months from 2015 to 2020 (SRC 2015–2020 spring; RMSE = 12.03 g C · kg$^{-1}$, $R^2$ = 0.64, RPIQ = 0.89), followed by the "SRC 2015–2020 spring_autumn" and the "SRC 2018–2020 full" bare soil composites with marginal differences (RMSE = 12.62 gC · kg$^{-1}$, $R^2$ = 0.63, RPIQ = 0.85 and RMSE = 12.57 gC · kg$^{-1}$, $R^2$ = 0.61, RPIQ = 0.85, respectively). These results suggest that SRC derived from images taken over a period of six years improves the accuracy of the prediction models. Moreover, the accuracies in the predictive performance of the single CNN models were slightly lower when only the reflectance values are used. This confirms the effectiveness of using multiple spectral channels as input to the CNN model.

The CNN model consistently outperformed the other machine learning models (PLS, RF, XGBooST and SVM). The predictive performances of all the models are provided in Table 7. It should be noted, that the Sentinel-2 absorbances used as input for all the models, since this pre-processing technique results in a better performance compared to the other spectral sources (reflectance and absorbance with SNV).

The SRC 2015–2020 spring composite has been used in the rest of the analysis since it attains the highest predictive performance. Fig. 7 illustrates the correlation of SOC content with the 10 Sentinel-2 bands as well as the environmental covariates. It should be noted that the highest correlations were observed in the VNIR regions.

Moreover, the efficacy of the local error correction layer in the CNN was evaluated. The overall performance, in terms of RMSE, and the $k$ neighbors of the CNN model that incorporates the local component are presented in Fig. 8. The localized correction layer was used to fine-tune the global CNN for each testing pattern separately adjusting its parameters by only considering the closet spectral neighbors of each unknown sample belonging to the calibration dataset and their accompanying prediction errors.

The outcomes show that the local adaptive correction technique has slightly improved SOC content prediction performance (RMSE = 11.95 gC · kg$^{-1}$, $R^2$ = 0.64, RPIQ = 0.90). The optimal $k$ neighbors are equal to 150.

Fig. 9 illustrates the regression plots of the measured against the predicted SOC values of the CNN model, the CNN version that incorporates the local learning approach and the DUPLICITE architecture.

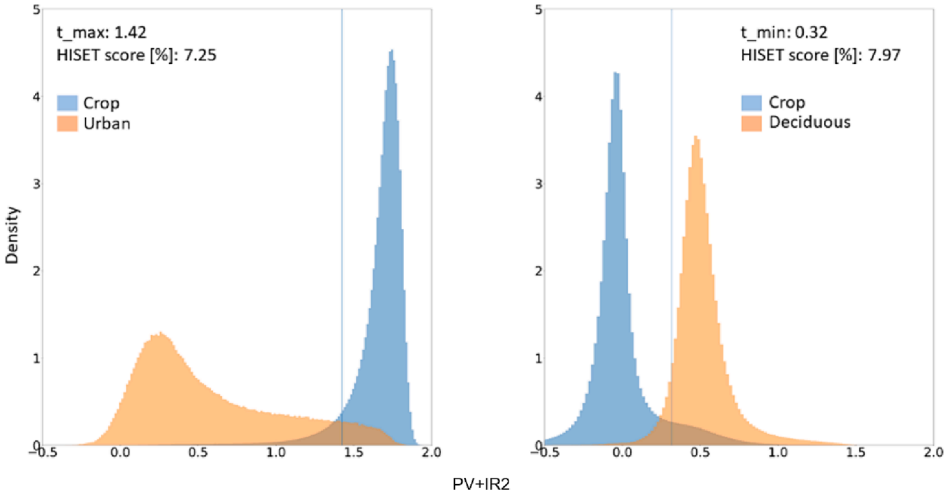



**Fig. 6.** a) Visualization of minimum and maximum thresholds for the different land uses as derived from PV + IR2 index.

**Table 6**
Statistical measures for evaluating the various SRCs across a range of seasons and time periods; in brackets, we present the number of LUCAS points covered.

| | SRC 2018–2020 full | SRC 2018–2020 spring_autumn | SRC 2018–2020 spring | SRC 2015–2020 spring_autumn | SRC 2015–2020 spring |
|---|---|---|---|---|---|
| Bare soil pixels (%) | 30.45 | 28.94 | 19.06 | 32.20 | 25.25 |
| No. of points (corresponding LUCAS indicated in brackets) covered by the SRC | 1650 (237) | 1616 (223) | 1050 (137) | 1714 (242) | 1417 (186) |
| **Spectral Angle (radians)** | | | | | |
| Mean | 0.084 | 0.080 | 0.069 | 0.082 | 0.075 |
| Median | 0.074 | 0.071 | 0.057 | 0.073 | 0.063 |
| St.dev. | 0.039 | 0.040 | 0.037 | 0.042 | 0.042 |

**Table 7**
Validation performances for SOC content estimation from the CNN and the competing ML models, across all the SRCs.

| | Models | PLS | RF | SVM | XGBooST | CNN | Single CNN |
|---|---|---|---|---|---|---|---|
| **Composite** | | | | | | | |
| SRC 2018–2020 full | RMSE | 15.07 | 13.22 | 15.01 | 13.61 | 12.57 | 12.66 |
| | $R^2$ | 0.46 | 0.58 | 0.59 | 0.57 | 0.61 | 0.61 |
| | RPIQ | 0.71 | 0.81 | 0.71 | 0.79 | 0.85 | 0.84 |
| SRC 2018–2020 spring_autumn | RMSE | 15.00 | 13.11 | 14.89 | 13.89 | 12.99 | 13.20 |
| | $R^2$ | 0.46 | 0.60 | 0.58 | 0.54 | 0.59 | 0.58 |
| | RPIQ | 0.71 | 0.82 | 0.72 | 0.77 | 0.82 | 0.81 |
| SRC 2018–2020 spring | RMSE | 14.89 | 13.48 | 14.35 | 14.11 | 12.65 | 12.51 |
| | $R^2$ | 0.44 | 0.54 | 0.55 | 0.51 | 0.60 | 0.61 |
| | RPIQ | 0.71 | 0.79 | 0.74 | 0.75 | 0.85 | 0.86 |
| SRC 2015–2020 spring_autumn | RMSE | 14.93 | 13.66 | 15.06 | 14.20 | 12.62 | 12.78 |
| | $R^2$ | 0.47 | 0.55 | 0.58 | 0.51 | 0.63 | 0.62 |
| | RPIQ | 0.72 | 0.78 | 0.71 | 0.75 | 0.85 | 0.84 |
| SRC 2015–2020 spring | RMSE | 14.64 | 12.93 | 14.29 | 13.92 | 12.03 | 12.36 |
| | $R^2$ | 0.48 | 0.58 | 0.57 | 0.52 | 0.64 | 0.62 |
| | RPIQ | 0.73 | 0.83 | 0.75 | 0.77 | 0.89 | 0.85 |

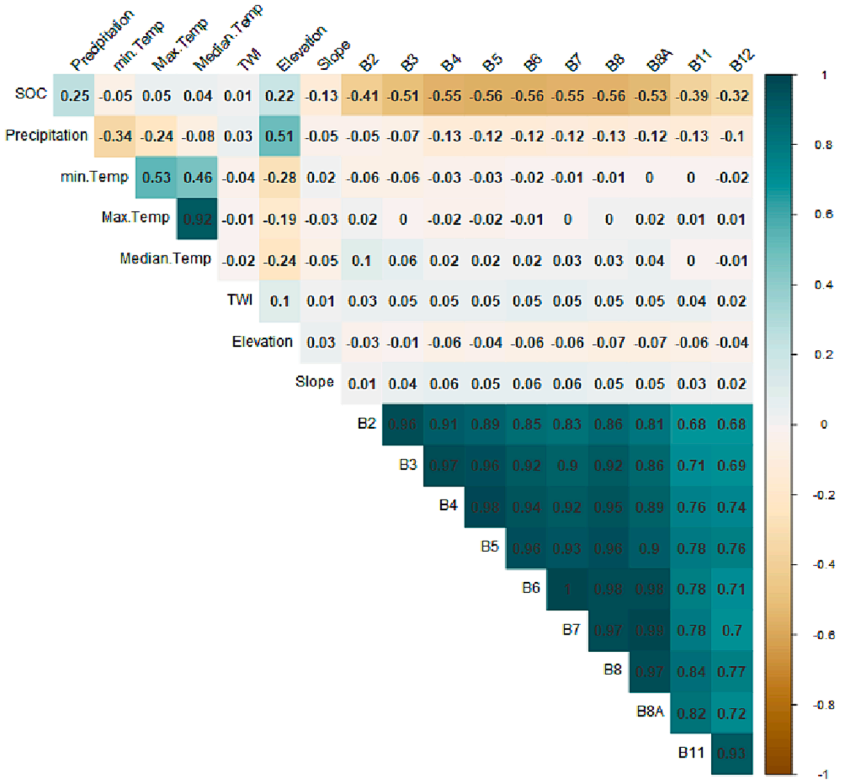



**Fig. 7.** Pearson's correlation for the SOC content from the selected topsoil sampling points along with the environmental covariates and the reflectance values obtained from the SRC 2015–2020 spring; Temp and TWI refer to temperature and topographic wetness index, respectively, while B refers to Sentinel-2 bands.

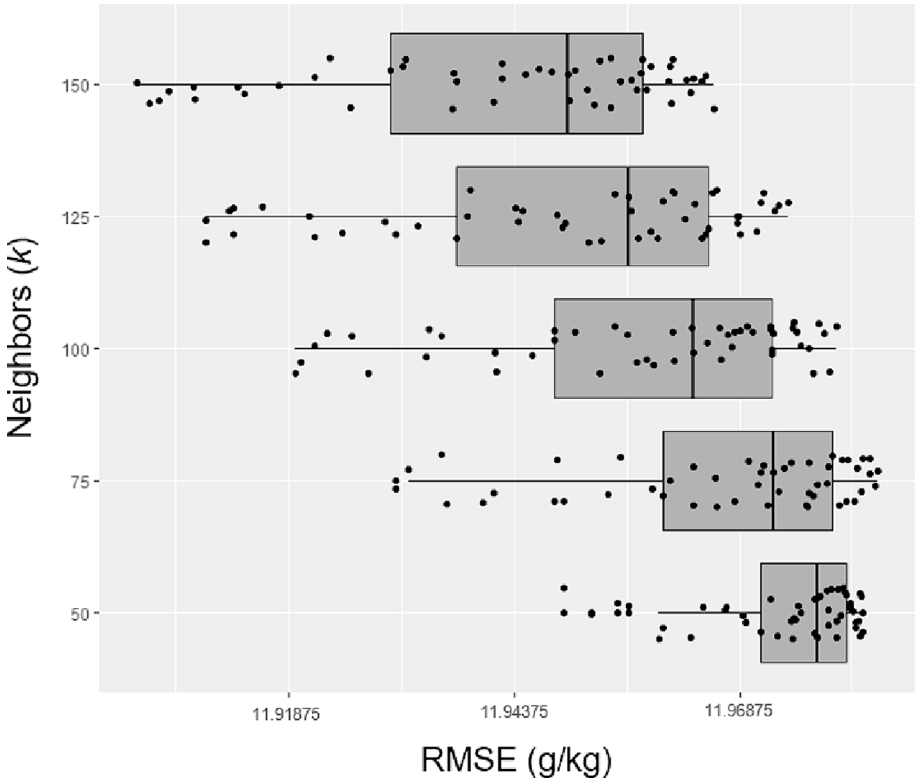



**Fig. 8.** Boxplots comparing the RMSEs generated from local-CNN by examining five different number of k neighbors.

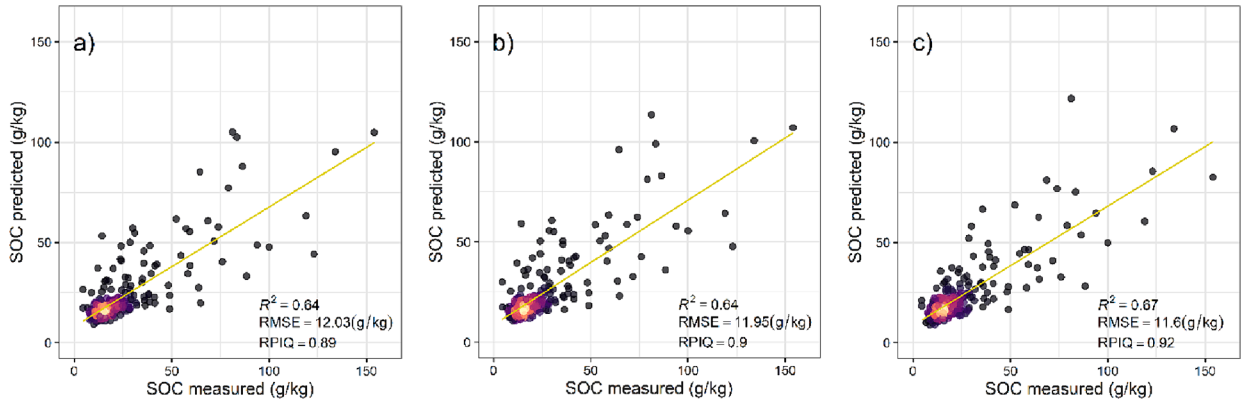



**Fig. 9.** Regression plots of measured against predicted SOC values considering the a) multi-input CNN architecture; b) the local-CNN that incorporates the local adaptive error correction mechanism and c) the novel DUPLICITE architecture. Color density has been utilized to illustrate the points.

The SRC 2015–202 spring composite has been considered as input in all the modelling activities. The scatter plots show that the predicted values produced from the CNN model have a low bias. Furthermore, the high SOC content values are underestimated because of the unevenly distributed SOC and the small number of high SOC values. Considering the performance, an improvement was achieved using DUPLICITE compared to CNN model that used only spectral inputs (RMSE = 11.06 g $C \cdot kg^{-1}$, $R^2 = 0.67$, RPIQ = 0.92). When comparing our model with the ML model proposed in the previous section, it becomes evident that all other models exhibit lower performance. In particular, we observed RMSE = 14.17 g $C \cdot kg^{-1}$, $R^2 = 0.52$, RPIQ = 0.76 for PLS, RMSE = 12.86 g $C \cdot kg^{-1}$, $R^2 = 0.62$, RPIQ = 0.83 for RF, RMSE = 14.31 g $C \cdot kg^{-1}$, $R^2 = 0.61$, RPIQ = 0.75 for SVM, and RMSE = 13.36 g $C \cdot kg^{-1}$, $R^2 = 0.57$, RPIQ = 0.80 for XGBooST.

### 3.3. Quality control and inspection of SOC map

Subsequently, the validated DUPLICITE model was applied to the "SRC 2015–2020 spring" composite. For our region of interest (about 30 % of the pixels have been identified as bare soil) the predictions seem free of any apparent artifacts and can be considered as visually homogenous (Fig. 10). Considering the entire Bavaria state, higher SOC predicted values can be observed in the Southeastern part, while SOC values in the west are higher compared to the predicted content in the eastern and central part of the Bavaria state.

Fig. 11 gives us a closer look at the SOC map product and allows us to investigate the changes and outcomes produced by the DUPLICITE model. Four test sites (Fig. 11) with representative agricultural areas have been selected. Moreover, it can be demonstrated how the proposed two branch architecture captured patterns of significant terrain heterogeneity, such as test site B and C. In conclusion, the visual assessment

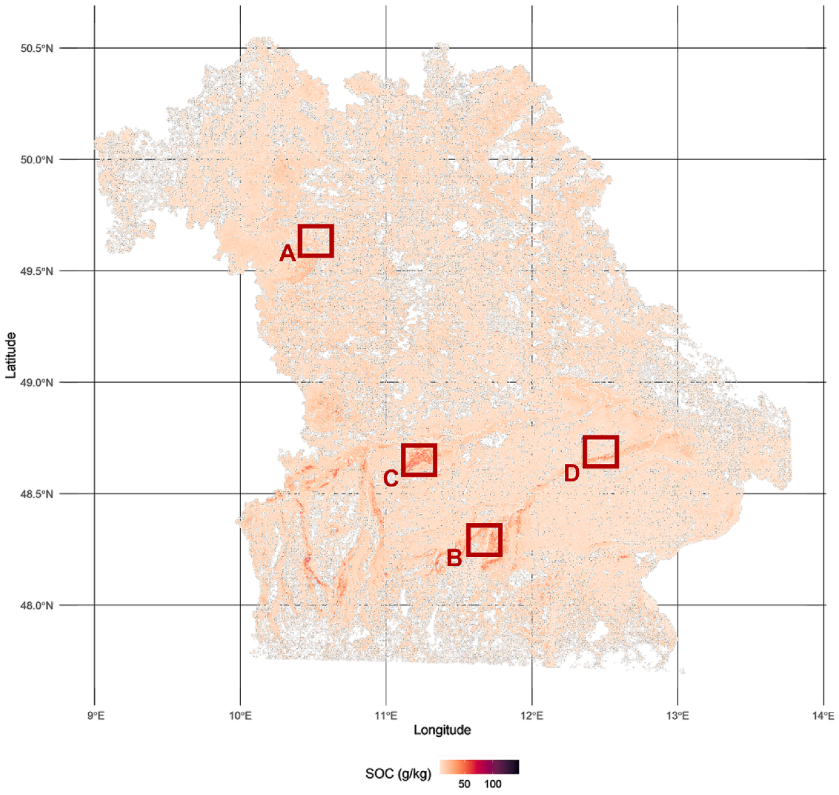



**Fig. 10.** SOC content spatial product as generated by the application of the DUPLICITE model based on the SRC 2015–2020 spring spatial data.

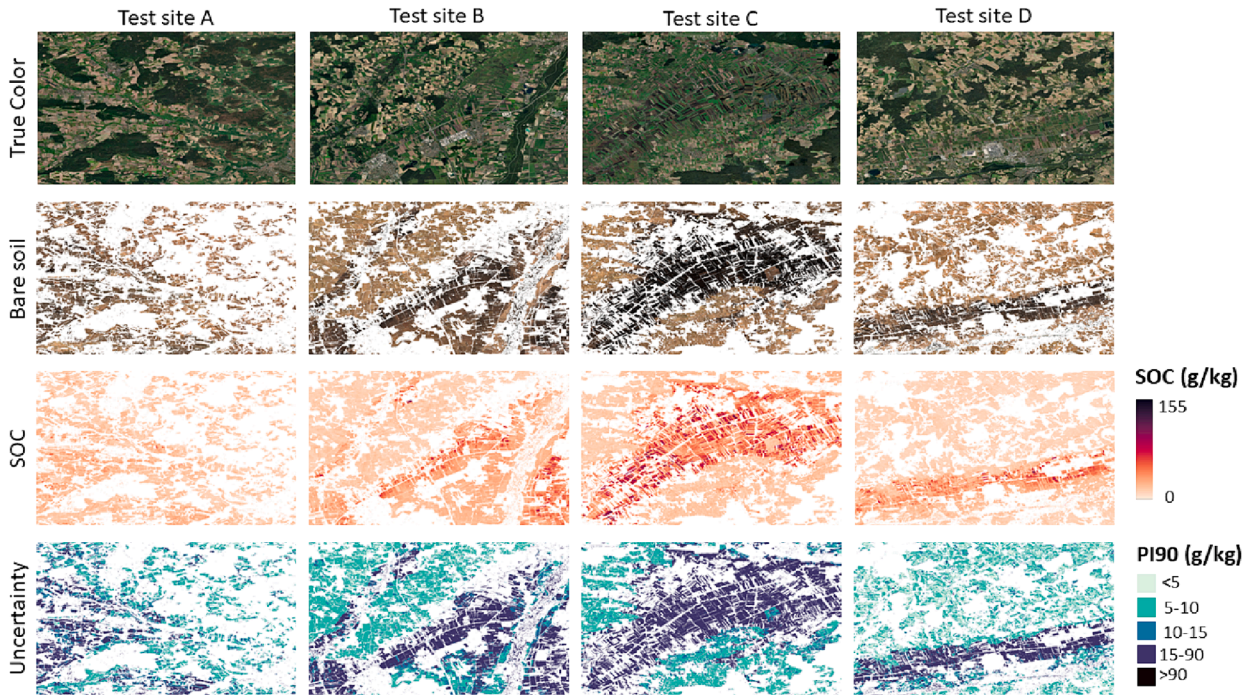



**Fig. 11.** Illustration of the true color images, the bare soil composites, the predicted SOC content maps, and the corresponding uncertainty products for representative test sites in Bavaria. Grassland and forest areas have masked out by white color.

is consistent with the quantitative findings (Section 3.2).

### 3.4. Interpretability analysis

We applied an interpretability analysis to explore which features are considered important for the estimation of the SOC content by each convolution layer. Fig. 12 depicts the most activated patterns (top) as well asl these that activated least (bottom) which generated from the filters comprising the 2nd convolution layer. For example, the patterns indicated that filters 2 and 3 use activations in the VNIR region to

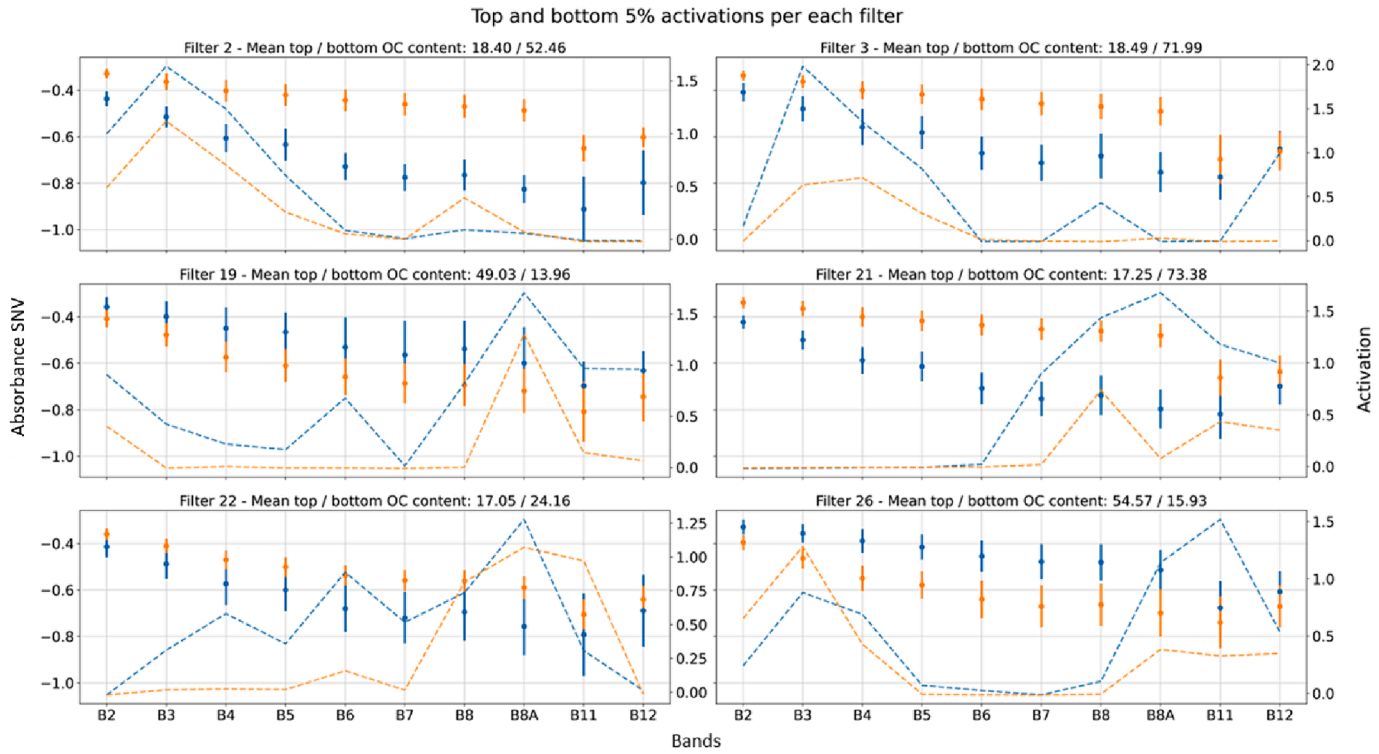



**Fig. 12.** Plots of the top 5% of activated patterns from the 2nd convolutional layer; corresponding to six filters chosen out of the total 32. The mean top (orange) and bottom (blue) activations are illustrated with the dashed lines. The orange and blue box plots represent the mean bottom and top Absorbance-SNV values.

effectively discriminate between the middle-values of medium (around Quantiles 2 and 3) and high SOC content. This is in agreement with the statement that the soil color in this spectral region is influenced by the presence of high SOC content values. The filters 19 and 21 identified medium–high and high SOC content values (3rd Quantile and beyond), while filter 22 can detect medium SOC content (i.e., values within the range of Quantiles 1 and 3). These are interesting findings, since it is proved that the CNN can generate complex features in both the VNIR and SWIR spectral regions and simultaneously discriminate the information contained in these regions in order to identify the SOC content. Furthermore, green band seems to activate both Filter 2 and Filter 3 to get information pertaining to the albedo that was affected by the SOC presence. The unique activations of Filter 26 in the SWIR region may be explained by the presence of loam and silt-loam soils in the test site.

## 4. Discussion

### 4.1. Soil reflectance composites assessment

Recent research has been motivated by the need to strike a balance between prediction performance in large-scale applications and adequate spatial coverage, without jeopardizing the selection of appropriate bare soil pixels (Dvorakova et al., 2021). It should be noted that the increased availability of bare soil pixels can prove to be advantageous by boosting the reliability and confidence of the SRC. Thus, we evaluated the potential of using independent thresholds per tile, to focus on bare soils by removing pixels with materials left on fields after the crop has been harvested and high soil moisture. Obviously, the SRC including spring month only showed the closest spectral similarity with the LUCAS spectra indicating lower contamination with disturbances. Further, the number of available model calibration points is essential for the subsequent SOC prediction, therefore, the more years are integrated, the better.

The main principles underpinning the proposed automated threshold derivation methodological framework are different from those of earlier methods which use ancillary data to distinguish areas between vegetated and bare states (Diek et al., 2017) or using pre-defined thresholds of spectral indices (Safanelli et al., 2021). The thresholds for the same indices can, however, vary slightly due to various characteristics occurring in each region. Thus, Castaldi (2021) proposed that the decision ought to be influenced by the needs of each research work in an effort to efficiently meet both the objectives of spatial coverage and prediction accuracy. The HISET approach may be accordingly applied to generate sets of thresholds for large diverse zones to fully automate the operational production of SRCs. To assure improved quality of the final SRCs, the threshold derivation may be implemented regionally in areas with significant changes in land cover and crop type.

In this paper, the use of the PV + IR2 index (Equation (1) allows us to minimize the potential false positives from pixels correspond to urban and non-photosynthetic vegetation areas. Vaudour et al. (2021) presented similar findings, where the bare soil composites generated using NBR2 and NDVI thresholds exhibited the best results for SOC content estimation, underlying the models' to both dry and green vegetation, in addition to soil moisture content. It should be taken in into account that NBR2 alone is not sufficient to mask soils which are covered with crop residues and are also wet (Dvorakova et al., 2020).

There is a marginal difference between the three-year composites including only seasonal months (spring to autumn and only spring) and those including all months (SRC 2018–2020 full, Table 6). The largest difference is observed between the "SRC 2018–2020 spring" composite and the "SRC 2018–2020 full" composite. As can be noted, the selection of only the spring months results in a much smaller covered area (ranging from 5 % to 10 %). In terms of spectrum similarity, the "SRC 2015–2020 spring_autumn" (0.082) is equivalent with "S2 2018–2020 full" (0.084). This may be attributed to the relatively wet spring and autumn seasons in our study area in 2017 that increased the occurrence of moist soil conditions.

However, the derived SOC predictions of all SRCs are marginal as discussed in detail below and are mainly driven by the number of available calibration points. It should be further noted, that especially in

Bavaria, winter crops are the dominant crop type and thus, also autumn month can provide reasonable bare soil pixels. Thus, the selection of the seasons for building SRCs should be decided based on the dominant crop type and on the seasonal weather conditions of the observed region.

### 4.2. Modelling discussion

#### 4.2.1. CNN impact analysis

DL techniques seems to be more effective in the prediction of soil properties via EO data (Odebiri et al., 2021; Yang et al., 2021) while CNN have been recognized as a powerful family of neural networks for prediction of SOC using various multi and hyperspectral data (Wang et al., 2022). Our results show that the evaluation metrics for all the SRCs utilizing the CNN do not greatly vary, but the six-year spring composite attains the best performance (RMSE = 12.03 gC · $kg^{-1}$; $R^2$ = 0.64; RPIQ = 0.89). In contrast to Wang et al. (2022), our findings indicated that CNN could achieve sufficient evaluation metrics, despite the limited spectral features (10 Sentinel-2 bands) used in convolution functions. Similarly, a significant improvement in the predictive performance obtained by the application of CNN models in the hyperspectral GaoFen-5 data (Meng et al., 2022). The use of multi-channel architecture utilizing the data derived from the pre-processing techniques demonstrates a notable advantage in generating an additional pool of spectral information from Sentinel-2 data. This allows us to yield a slight enhancement in the predictions of the CNN model ($RMSE_{mean}$ = 0.15 gC · $kg^{-1}$) as presented in Table 7. It should be noted that, the spectral space generating by the variations of the hyperspectral transformations shows overwhelming advantages compared to multispectral data (Tsimpouris et al., 2021).

We can notice an improvement in the evaluation metrics by keeping only the spring months when the likelihood of bare soils is highest (e.g., seedbed condition). They indicated that the optimal time for bare soil selection matches the last observation before the crop develops. A previous work by Vaudour et al. (2021) also highlighted that spring period is more suited to generate bare soil composites compared to autumn and winter months, due to the influence of solar elevation in the Northern hemisphere. In that regard, the performance can be slightly enhanced by taking into account images extending over a longer period (three vs. six years SRCs, Table 7). Here, we also demonstrate that the CNN performance can be improved by incorporating information derived from a local spectral neighborhood into an adaptive error-correction mechanism Fig. 8.

A potential shortcoming of DL techniques is their low interpretability. Overall, limited studies address this issue by generating uncertainty spatial products along with predictive maps for the target properties (Dvorakova et al., 2023). In this context, we address the interpretability issue, by exploiting summarization approaches aggregating pattern activations and by generating attribution graphs indicating the most important spectral features. As a result, we are in a position to understand which features are considered important for the estimation of the SOC content by each convolution layer, highlighting the CNN's ability to generate features in both VNIR and SWIR spectral regions, enabling simultaneous differentiation and identification of SOC information within these regions.

#### 4.2.2. DUPLICITE impact analysis

In our work, the proposed DUPLICITE architecture puts forward the combination of i) the optical data processed by the first branch (CNN) ii) with the supplementary information derived from environmental covariates processed by the second branch (ANN). The efficient fusion of these two branches results in significantly improved prediction accuracies when compared to conventional ML models (Table 7) and alternative CNN frameworks that are based on the spectral information (Fig. 9). Recently, Zeraatpisheh et al. (2022) presented the idea of combining diverse EO sources. However, the novelty introduced by DUPLICITE is the way to fuse these supplementary EO-based information. Despite the fact that neither CNN performs well on its own (RMSE = 12.03 gC · $kg^{-1}$; $R^2$ = 0.64; RPIQ = 0.89), the accuracy slightly improves when used in conjunction with ANN to manage the environmental and terrain covariates (RMSE = 11.60 gC · $kg^{-1}$; $R^2$ = 0.67; RPIQ = 0.92). The better predictive performance achieved by DUPLICITE may be explained due to the correlation of some of the environmental covariates with SOC content (Fig. 7). DUPLICITE provides an end-to-end framework towards addressing the complexity of different EO data streams (e.g., optical data and environmental covariates) for topsoil mapping (Fig. 10).

### 4.3. Limitations and outlook

Despite the encouraging results presented in Table 7, there are some limitations. First, it should be noted that the few spectral bands of Sentinel-2 are the most significant hurdle for estimating SOC with higher accuracy, a limitation that may be lifted by the forthcoming hyperspectral missions. Mzid et al., (2022), as revealed in their recent study, demonstrated hyperspectral PRISMA satellite outperformed Sentinel-2 in achieving the highest accuracy in retrieving topsoil SOC.

It should be noted that it is still difficult to distinguish bare soil pixels from multispectral satellite imagery data. Therefore, the indices' choice is not a panacea and there does not exist a single optimal value for the thresholds that can be used globally. Additional indices should be evaluated considering the multispectral systems that are limited to coarse band in the SWIR region. The future superspectral Landsat and Sentinel generations (Wu et al., 2019) and the imaging spectrometers that will incorporate more channels in the SWIR can be beneficial for non-photosynthetic vegetation monitoring (Dennison et al., 2023). Similarly, Yue et al. (2020) introduced the Broadband Spectral Angle Index to alleviate the impact on spectral reflectance of plant residues and soil moisture content, which helps in the selection of more representative pixels. In a recent work, Tian et al. (2023) proposed a VNIR spectral shape index for bare soil discrimination to fully utilize current and historical satellite systems.

Notwithstanding the improvements resulting from our proposed CNNs, Fig. 9 illustrates underestimation of high values. Class imbalance is common in soil datasets, raising a need for techniques to improve model's attention to underrepresented classes to improve predictive performance. Recent experiments have tested CNN architectures incorporating attention mechanisms (Pu et al., 2021), while others have combined them with a synthetic minority oversampling technique for further optimization (Hamdi et al., 2022).

In AI-based regression algorithms the size of the dataset (both in terms of feature space and number of labelled patterns) as well as the variance exhibited within is a further crucial factor (Padarian et al., 2020). Therefore, it is imperative to assess and evaluate the capabilities of the DL models on larger datasets, such as LUCAS (Tóth et al., 2013) and the dataset from the Rapid Carbon Assessment Project in the United States (Wijewardane et al., 2016). This will enable a thorough examination of its effectiveness in large-scale applications when compared to other ML models. The proposed CNN may be further tweaked and optimized by automatically selecting its hyperparameters and network topology leading to optimal frameworks (Shen and Viscarra Rossel, 2021). In a broader perspective, the proposed models, i.e., the CNN architecture of Fig. 3 and the dual input configuration of Fig. 5, can be readily extended to other multispectral and recently launched hyperspectral space-borne sensors and to estimate other soil properties (e.g., pH, $CaCO_3$, etc.).

Further, the proposed CNN architecture may be extended through the integration of other types of EO data. For instance, Sentinel-1 radar time series data may be incorporated, to diminish the effect of ambient factors in optical data. In this context, the inclusion of Sentinel-1 derived moisture maps to facilitate modeling of SOC croplands has been presented by Urbina-Salazar et al. (2021), while (Mirzaeitalarposhti et al., 2022) used different combinations of EO data (i.e., optical, radar and

environmental covariates) to estimate soil texture fractions.

Tziolas et al. (2020) introduced a multi-input DL framework based on CNN to estimate clay content leveraging the information contained in Sentinel-1 and Sentinel-2 time series data. On step forward, Interdonato et al. (2019) suggested the synergistic use of a CNN to generate deeper information from spatial autocorrelation, with recurrent networks to explore the temporal dependencies. This rationale can be adopted to enhance the DUPLICITE model by involving the application of such different yet complementary architectures. A hybrid regression framework may result in a more accurate representation of the distortions in the optical data incurred by soil moisture and roughness. For instance, Ienco et al. (2019) proposed a model which utilizes the deep features extracted through a CNN and then uses a RF model to infer the final output. The new spaceborne hyperspectral data (Meng et al., 2022), as well as an ever-expanding set of environmental covariates may be considered (Hengl et al., 2021; Saidi et al., 2022) to enhance the accuracy of the DUPLICITE framework. Especially for environmental covariates CNN architectures may be enhanced by integrating 2D convolutional filters to better exploit the spatial autocorrelation between neighboring pixels. Also, a possible fusion of multispectral images obtained from different satellites can also be evaluated to achieve more detailed information as opposed to a single sensor.

## 5. Conclusions

For the purpose of predicting SOC content over croplands, we demonstrated how DL methods and soil reflectance composites offer tremendous potential. The current work aspires to spark a number of new promising research studies to address the unprecedented challenges in soil mapping leveraging and build upon the multi-temporal spaceborne data.

We demonstrated that three-year time frame is adequate to create multidate mosaics with sufficient bare soil coverage and simultaneously yield satisfactory prediction performance for SOC content. Overall, a SRC spanning a six year period is more favored, however the selection of the months to include has a significant impact (i.e., preferably spring months). The importance of our proposal has been proved by a multi-channel CNN evaluation over the Bavaria state, demonstrating how it outperforms current state-of-the-art AI models. We also concluded that the complimentary data resulting from different spectral transformations could enhance the overall accuracy.

The DUPLICITE architecture is consisted of discrete neural networks able to process both spectral data (e.g., Sentinel-2) and additional covariates (e.g., environmental and terrain) enforcing a diverse representation of the available information. The final model concatenates the features extracted by each branch. It should be noted that an improvement of 3.57 % in RMSE has been achieved compared to the results of the CNN that uses only spectral information.

## 6. Availability of material

The produced prediction map of SOC content using the CNN model is available as Cloud-Optimized GeoTIFF in static.i-bec.org/worldsoils/SOC.Bavaria.html.

## Funding

The research leading to these findings and results has been implemented during the WORLDSOILS project that has received funding from the European Space Agency within the framework of the EO Science for Society slice of the 5th Earth Observation Envelope Program.

## CRediT authorship contribution statement

**Nikolaos Tziolas:** Writing – original draft, Methodology. **Nikolaos Tsakiridis:** Writing – original draft, Methodology. **Uta Heiden:** Validation, Methodology, Data curation. **Bas van Wesemael:** Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

The authors would like to thank the Bavarian State Research Center for Agriculture (LFL) and the Bavarian Environment Agency (LFU) for providing the soil datasets.

## References

Asgari, N., Ayoubi, S., Jafari, A., Demattê, J.A.M., 2020. Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. Int. J. Remote Sens. 41, 7624–7648. https://doi.org/10.1080/01431161.2020.1763506.

Azizi, K., Garosi, Y., Ayoubi, S., Tajik, S., 2023. Integration of Sentinel-1/2 and topographic attributes to predict the spatial distribution of soil texture fractions in some agricultural soils of western Iran. Soil Tillage Res. 229, 105681 https://doi.org/10.1016/j.still.2023.105681.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrol. Sci. Bull. 24, 43–69. https://doi.org/10.1080/02626667909491834.

Breiman, L., 2001. Random forests. Mach. Learn. https://doi.org/10.1023/A:1010933404324.

Bruneau, P., McElroy, N.R., 2006. logD7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. J. Chem. Inf. Model. 46, 1379–1387. https://doi.org/10.1021/ci0504014.

Castaldi, F., 2021. Sentinel-2 and Landsat-8 multi-temporal series to estimate topsoil properties on croplands. Remote Sens. (Basel). https://doi.org/10.3390/rs13173345.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system, in. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Dennison, P.E., Lamb, B.T., Campbell, M.J., Kokaly, R.F., Hively, W.D., Vermote, E., Dabney, P., Serbin, G., Quemada, M., Daughtry, C.S.T., Masek, J., Wu, Z., 2023. Modeling global indices for estimating non-photosynthetic vegetation cover. Remote Sens. Environ. 295, 113715 https://doi.org/10.1016/j.rse.2023.113715.

Diek, S., Fornallaz, F., Schaepman, M.E., de Jong, R., 2017. Barest pixel composite for agricultural areas using landsat time series. Remote Sens. (Basel). https://doi.org/10.3390/rs9121245.

Drucker, H., Surges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines, in: Advances in Neural Information Processing Systems.

Dvorakova, K., Shi, P., Limbourg, Q., van Wesemael, B., 2020. Soil organic carbon mapping from remote sensing: the effect of crop residues. Remote Sens. (Basel). https://doi.org/10.3390/rs12121913.

Dvorakova, K., Heiden, U., van Wesemael, B., 2021. Sentinel-2 exposed soil composite for soil organic carbon prediction. Remote Sens. (Basel). https://doi.org/10.3390/rs13091791.

Dvorakova, K., Heiden, U., Pepers, K., Staats, G., van Os, G., van Wesemael, B., 2023. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. Geoderma 429, 116128. https://doi.org/10.1016/j.geoderma.2022.116128.

Gasmi, A., Gomez, C., Lagacherie, P., Zouari, H., Laamrani, A., Chehbouni, A., 2021. Mean spectral reflectance from bare soil pixels along a landsat-TM time series to increase both the prediction accuracy of soil clay content and mapping coverage. Geoderma 388, 114864. https://doi.org/10.1016/j.geoderma.2020.114864.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Hagolle, O., Huc, M., Desjardins, C., Auer, S., R.Richter, 2017. MAJA Algorithm Theoretical Basis Document (1.0) [WWW Document]. Zenodo. 10.5281/zenodo.1209633.

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755. https://doi.org/10.1016/j.rse.2010.03.002.

Hamdi, S., Oussalah, M., Moussaoui, A., Saidi, M., 2022. Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound. J. Intell. Inf. Syst. 59, 367–389. https://doi.org/10.1007/s10844-022-00707-7.

Heiden, U., d'Angelo, P., Schwind, P., Karlshöfer, P., Müller, R., Zepp, S., Wiesmeier, M., Reinartz, P., 2022. Soil reflectance composites—improved thresholding and performance evaluation. Remote Sens. (basel). https://doi.org/10.3390/rs14184526.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. https://doi.org/10.1038/s41598-021-85639-y.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorolog. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Ienco, D., Interdonato, R., Gaetano, R., Minh, H.T., D.,, 2019. Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2019.09.016.

Interdonato, R., Ienco, D., Gaetano, R., Ose, K., 2019. DuPLO: a DUal view point deep Learning architecture for time series classificatiOn. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2019.01.011.

ISO 10694:1995 Soil Quality Determiantion of Organic and Total Carbon after Dry Compustion (Elementary Analysis). International Organization for Standardization, Geneva.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Kühnel, A., Wiesmeier, M., Kögel-Knabner, I., Spörlein, P., 2020. Veränderungen der Humusqualität und Quantität Bayerischer Böden im Klimawandel.

Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Bourennane, H., Saby, N.P.A., Martin, M.P., Vaudour, E., Gomez, C., Lagacherie, P., Arrouays, D., 2019. Satellite data integration for soil clay content modelling at a national scale. Int. J. Appl. Earth Obs. Geoinf. 82, 101905 https://doi.org/10.1016/j.jag.2019.101905.

Meng, X., Bao, Y., Wang, Y., Zhang, X., Liu, H., 2022. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. Remote Sens. Environ. 280, 113166 https://doi.org/10.1016/j.rse.2022.113166.

Minasny, B., McBratney, A.B., 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.

Mirzaeitalarposhti, R., Shafizadeh-Moghadam, H., Taghizadeh-Mehrjardi, R., Demyan, M.S., 2022. Digital soil texture mapping and spatial transferability of machine Learning models using Sentinel-1, Sentinel-2, and terrain-derived covariates. Remote Sens. (Basel). https://doi.org/10.3390/rs14235909.

Möller, M., Zepp, S., Wiesmeier, M., Gerighausen, H., Heiden, U., 2022. Scale-specific prediction of topsoil organic carbon contents using terrain attributes and SCMaP soil reflectance composites. Remote Sens. (Basel). https://doi.org/10.3390/rs14102295.

Montanarella, L., 2015. Agricultural policy: govern our soils. Nature 528, 32–33. https://doi.org/10.1038/528032a.

Mzid, N., Pignatti, S., Huang, W., Casa, R., 2021. An analysis of Bare soil occurrence in arable croplands for remote sensing topsoil applications. Remote Sens. (Basel). https://doi.org/10.3390/rs13030474.

Mzid, N., Castaldi, F., Tolomio, M., Pascucci, S., Casa, R., Pignatti, S., 2022. Evaluation of agricultural Bare soil properties retrieval from landsat 8, Sentinel-2 and PRISMA satellite data. Remote Sens. (Basel). https://doi.org/10.3390/rs14030714.

Odebiri, O., Odindi, J., Mutanga, O., 2021. Basic and deep learning models in remote sensing of soil organic carbon estimation: a brief review. Int. J. Appl. Earth Obs. Geoinf. 102, 102389 https://doi.org/10.1016/j.jag.2021.102389.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153. https://doi.org/10.1111/ejss.12499.

Padarian, J., Minasny, B., McBratney, A.B., 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.

Paul, C., Bartkowski, B., Dönmez, C., Don, A., Mayer, S., Steffens, M., Weigl, S., Wiesmeier, M., Wolf, A., Helming, K., 2023. Carbon farming: are soil carbon certificates a suitable tool for climate change mitigation? J. Environ. Manage. 330, 117142 https://doi.org/10.1016/j.jenvman.2022.117142.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. SOIL. 10.5194/soil-7-217-2021.

Pu, C., Huang, H., Yang, L., 2021. An attention-driven convolutional neural network-based multi-level spectral–spatial feature learning for hyperspectral image classification[formula presented]. Expert Syst. Appl. 185 https://doi.org/10.1016/j.eswa.2021.115663.

Rizzo, R., Wadoux, A.M.J.-C., Demattê, J.A.M., Minasny, B., Barrón, V., Ben-Dor, E., Francos, N., Savin, I., Poppiel, R., Silvero, N.E.Q., Terra, F. da S., Rosin, N.A., Rosas, J.T.F., Greschuk, L.T., Ballester, M.V.R., Gómez, A.M.R., Belllinaso, H., Safanelli, J. L., Chabrillat, S., Fiorio, P.R., Das, B.S., Malone, B.P., Zalidis, G., Tziolas, N., Tsakiridis, N., Karyotis, K., Samarinas, N., Kalopesa, E., Gholizadeh, A., Shepherd, K. D., Milewski, R., Vaudour, E., Wang, C., Salama, E.S.M., 2023. Remote sensing of the Earth's soil color in space and time. Remote Sens Environ 299, 113845. 10.1016/j.rse.2023.113845.

Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T., Heiden, U., 2018. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with landsat imagery (1984–2014). Remote Sens. Environ. 205, 1–17. https://doi.org/10.1016/j.rse.2017.11.004.

Safanelli, J.L., Demattê, J.A.M., Chabrillat, S., Poppiel, R.R., Rizzo, R., Dotto, A.C., Silvero, N.E.Q., Mendes, W. de S., Bonfatti, B.R., Ruiz, L.F.C., ten Caten, A., Dalmolin, R.S.D., 2021. Leveraging the application of Earth observation data for mapping cropland soils in Brazil. Geoderma. 10.1016/j.geoderma.2021.115042.

Saidi, S., Ayoubi, S., Shirvani, M., Azizi, K., Zeraatpisheh, M., 2022. Comparison of different machine learning methods for predicting cation exchange capacity using environmental and remote sensing data. Sensors. https://doi.org/10.3390/s22186890.

Shen, Z., Viscarra Rossel, R.A., 2021. Automated spectroscopic modelling with optimised convolutional neural networks. Sci. Rep. 11, 208. https://doi.org/10.1038/s41598-020-80486-9.

Tetko, I.V., Tanchuk, V.Y., 2002. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J. Chem. Inf. Comput. Sci. 42, 1136–1145. https://doi.org/10.1021/ci025515j.

Tian, J., Zhang, Z., Philpot, W.D., Tian, Q., Zhan, W., Xi, Y., Wang, X., Zhu, C., 2023. Simultaneous estimation of fractional cover of photosynthetic and non-photosynthetic vegetation using visible-near infrared satellite imagery. Remote Sens. Environ. 290, 113549 https://doi.org/10.1016/j.rse.2023.113549.

Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185, 7409–7425. https://doi.org/10.1007/s10661-013-3109-3.

Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma. https://doi.org/10.1016/j.geoderma.2020.114208.

Tsimpouris, E., Tsakiridis, N.L., Theocharis, J.B., 2021. Using autoencoders to compress soil VNIR–SWIR spectra for more robust prediction of soil properties. Geoderma. https://doi.org/10.1016/j.geoderma.2021.114967.

Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., Zalidis, G., 2020. Employing a multi-input deep convolutional neural network to derive soil clay content from a synergy of multi-temporal optical and radar imagery data. Remote Sens. (Basel). https://doi.org/10.3390/rs12091389.

Tziolas, N., Tsakiridis, N., Chabrillat, S., Demattê, J.A.M., Ben-Dor, E., Gholizadeh, A., Zalidis, G., van Wesemael, B., 2021. Earth observation data-driven cropland soil monitoring: a review. Remote Sens. (Basel). https://doi.org/10.3390/rs13214439.

Urbina-Salazar, D., Vaudour, E., Baghdadi, N., Ceschia, E., Richer-de-Forges, A.C., Lehmann, S., Arrouays, D., 2021. 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 Sensing . 10.3390/rs13245115.

Vaudour, E., Gomez, C., Lagacherie, P., Loiseau, T., Baghdadi, N., Urbina-Salazar, D., Loubet, B., Arrouays, D., 2021. Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands. Int. J. Appl. Earth Obs. Geoinf. https://doi.org/10.1016/j.jag.2020.102277.

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the landsat 8/OLI land surface reflectance product. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2016.04.008.

Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A.J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., Huang, Y., 2022. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: assessing potential of airborne and spaceborne optical soil sensing. Remote Sens. Environ. 271, 112914 https://doi.org/10.1016/j.rse.2022.112914.

Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., Schilling, B., von Lützow, M., Kögel-Knabner, I., 2012. Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth. Glob. Chang. Biol. 18, 2233–2245. https://doi.org/10.1111/j.1365-2486.2012.02699.x.

Wiesmeier, M., Mayer, S., Burmeister, J., Hübner, R., Kögel-Knabner, I., 2020. Feasibility of the 4 per 1000 initiative in Bavaria: a reality check of agricultural soil management and carbon sequestration scenarios. Geoderma 369, 114333. https://doi.org/10.1016/j.geoderma.2020.114333.

Wijewardane, N.K., Ge, Y., Wills, S., Loecke, T., 2016. Prediction of soil carbon in the conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. Soil Sci. Soc. Am. J. 80, 973–982. https://doi.org/10.2136/sssaj2016.02.0052.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intel. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Wu, Z., Snyder, G., Vadnais, C., Arora, R., Babcock, M., Stensaas, G., Doucette, P., Newman, T., 2019. User needs for future landsat missions. Remote Sens. Environ. 231, 111214 https://doi.org/10.1016/j.rse.2019.111214.

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. Int. J. Appl. Earth Obs. Geoinf. 102, 102428 https://doi.org/10.1016/j.jag.2021.102428.

Yue, J., Tian, Q., Dong, X., Xu, N., 2020. Using broadband crop residue angle index to estimate the fractional cover of vegetation, crop residue, and bare soil in cropland

systems. Remote Sens. Environ. 237, 111538 https://doi.org/10.1016/j.rse.2019.111538.
Zepp, S., Heiden, U., Bachmann, M., Wiesmeier, M., Steininger, M., van Wesemael, B., 2021. Estimation of soil organic carbon contents in croplands of Bavaria from SCMaP soil reflectance composites. Remote Sens. (Basel). https://doi.org/10.3390/rs13163141.
Zeraatpisheh, M., Garosi, Y., Reza Owliaie, H., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. Catena (amst) 208, 105723. https://doi.org/10.1016/j.catena.2021.105723.
Žížala, D., Minarík, R., Zádorová, T., 2019. Soil organic carbon mapping using multispectral remote sensing data: prediction ability of data with different spatial and spectral resolutions. Remote Sens. (Basel). https://doi.org/10.3390/rs11242947.