ISPRS Journal of Photogrammetry and Remote Sensing 213 (2024) 72–86

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Evaluating the spatial–temporal transferability of models for agricultural land cover mapping using Landsat archive

Jayan Wijesingha *, Ilze Dzene, Michael Wachendorf

Grassland Science and Renewable Plant Resources, Universität Kassel, Steinstraße 19, D-37213, Witzenhausen, Germany

ARTICLE INFO

Keywords:
Agricultural land cover
Crop types
Landsat
Spatial–temporal transferability
Machine learning

ABSTRACT

Changes in policy and new plans can significantly influence land use and trigger land use change in the long term. The data for pre- and post-policy implementation is necessary to assess the specific policy's impact on land use. In the early nineties, Germany started promoting renewable energy production, including bioenergy, which changed the agricultural landscape. Remote sensing (RS) image-based machine learning models can be beneficial for mapping agricultural land use in the present and the past. However, machine learning classification models trained on RS data from specific training sites and time may not be able to predict data for unknown sites and unknown temporal points due to changes in crop phenology, field features, or ecological site circumstances because most of the models are limited in their performances according to variations of the training data set. Therefore, this study aims to assess the spatial–temporal transferability of Landsat-based agricultural land use type classification. The study was developed to map agricultural land cover (5 classes: maize, grasslands, summer crops, winter crops, and mixed crops) in two regions in Germany (North Hesse and Weser-Ems) between the years 2010 and 2018 using Landsat archive data (i.e., Landsat 5, 7, and 8). Two machine learning models (random forest – RF and 2D convolution neural network – 2DCNN) were trained and evaluated according to no transferability (reference) scenario and three spatial–temporal scenarios using mF1 and class level F1 values. Three model transferability scenarios were evaluated: a) temporal – S1, b) spatial – S2, and c) spatiotemporal – S3. The reference scenario, without transferability, achieved an overall accuracy of 89.1% and a macro F1 score of 0.74 for RF and 89.9% and 0.75 for CNN, respectively. Under three transferability scenarios (S1, S2, and S3), the macro F1 scores decreased to 0.67, 0.66, and 0.62 for RF, and 0.68, 0.62, and 0.58 for CNN, respectively. The dissimilarity between the data employed to train the model and data from the new domain indicated a clear link that could explain the reduction in model predictability. Moreover, the performance degradation could be attributed to the disparity in environmental, climatic, and crop calendar conditions between the two domains. Understanding the extent of model performance degradation during transferability is crucial for developing effective strategies to mitigate these issues and enhance the generalisability of machine learning models for agriculture land cover mapping.

1. Introduction

During the last century, the rapid growth of the population has caused changes in the Earth's land use and land cover (LULC) at an alarming rate (Hooke et al., 2012). The land cover change represents an alteration of the cover, and land use change indicates a modification of how the land is being utilised or handled. The combination of both processes is called land use and land cover change (LULCC). Anthropogenic activities to support human needs (e.g., for food, fibre, and energy) are the primary cause of the LULCC.

Multiple reports and studies have shown that in the last century, most of the natural land cover (e.g., primary forests, natural grasslands) has been gradually converted to arable lands to grow crops. According to the FAO data, the total agricultural land cover in the world increased by

Abbreviations: AOA, area of applicability; CNN, convolution neural network; DA, domain adaptation; DI, dissimilarity index; DL, deep learning; EVI, enhanced vegetation index; IACS, Integrated Administration and Control System; LULC, land use and land cover; LULCC, land use and land cover change; ML, machine learning; NDMI, normalised difference moisture index; NDVI, normalised difference vegetation index; NH, North Hesse; OA, overall accuracy; OSAVI, optimized soil adjusted vegetation index; RF, random forest; RS, remote sensing; SGD, stochastic gradient descent; VI, vegetation index; WE, Weser-Ems.

* Corresponding author.
E-mail address: jayan.wijesingha@uni-kassel.de (J. Wijesingha).

https://doi.org/10.1016/j.isprsjprs.2024.05.020
Received 29 August 2023; Received in revised form 15 May 2024; Accepted 22 May 2024
Available online 31 May 2024

six percent between 1961 and 2019, accounting for 282 million hectares (FAO, 2023). Apart from food and feed production, the utilisation of cropland for non-food products and services also triggered the expansion of agricultural land cover.

Understanding the status of the agricultural land cover (e.g., in terms of the total cropland area and distribution of crop types) and its change over the spatial–temporal domain is vital for food security and sustainable development (Pérez-Hoyos et al., 2017; See et al., 2015). Moreover, for many countries and regions, remote sensing (RS) studies are the best, or only, way to obtain detailed information on historical land-use conditions (Braun, 2021). RS based methods (Asam et al., 2022; Blickensdörfer et al., 2022; Ofori-Ampofo et al., 2021) have shown the potential to acquire efficient and accurate information about agricultural land cover. The availability of time series RS satellite data (both optical and RADAR) and the application of machine learning (ML) algorithms boosted large-scale cropland mapping (Asam et al., 2022; Ofori-Ampofo et al., 2021).

Most of the published cropland mapping studies are using freely available medium-resolution satellite image time series such as Sentinel 1 & 2 and Landsat with supervised ML models. For example, the recent study by Asam et al. (2022) used Sentinel 1 & 2 time series data with random forest (RF) ML algorithm for mapping 17 crop types in Germany for the year 2018 at 10 m spatial resolution with 75.5 % overall accuracy. However, the Sentinel data with 10 m spatial resolution are available only since 2015. To evaluate the long-term impact of policies encouraging potentially destructive land use activities, time-series RS data covering broader time horizons are vital (Song, 2023). For example, the Renewable Energy Act introduced in Germany in year 2000 was intended to promote energy generation from renewable sources. Electricity generated in biogas plants using dedicated energy crops (e.g., maize silage) as feedstock were particularly supported. In the following years, the change of agricultural land cover and land use in many parts of Germany were observed, raising concerns and discussions about "maizifaction" of the landscape for biogas production (Vergara and Lakes, 2019). In this particular case, covering the whole period of the policy impact using RS data requires information about land cover and land use going back until the year 2000. To obtain cropland data and its spatial distribution before the year 2015, the Landsat data with 30 m spatial resolution can be employed (Kyere et al., 2019). Landsat provides RS data covering the time since the early 1970 s until nowadays using a series of satellites (i.e., Landsat 1–9). Kyere et al. (2019) combined Landsat data and cropland parcel boundaries from the states to map four crop types in the North Hesse region in Germany and achieved 71 % overall accuracy by using the RF model.

Many studies on cropland or crop-type mapping with RS satellite data are using field data from the same year to train and validate the models (Wu et al., 2023). This is due to the limitation of the field data availability in other years. Besides the lack of training data from different years, the performance of the models is further challenged by changes in crop phenology due to annual weather patterns and farming practices. Most of the created models are good at predicting data in the same domain as the training data but are decreasing their performance when transferred to other temporal contexts. To avoid the limitation of model transferability in different periods, Kyere et al. (2019) suggested employing multi-year data to train models. By including multi-year data for the training phase, the model learns different annual crop phenology patterns that help improving predictions for the other years. As reported by Kyere et al. (2019), a six percent increment in overall accuracy was observed when the model was trained using multi-year data.

Similarly, studies reporting cropland mapping with RS data are mostly limited to a certain geographical region. When the model developed and trained for one region is applied to predict values in the different spatial regions, the accuracy of predicted values might be reduced due to significant spatial change. However, if the model has been trained with data from multiple spatial regions, it can be effectively transferred to a new spatial region (Orynbaikyzy et al., 2022). A recent study by Ajadi et al., (2021) showed that cropland mapping with multi-region training data using the XGBoost ML model can be accurately transferred to another spatial region. Similarly, Orynbaikyzy et al. (2022) reported that using multi-state training data with the RF model could help to reduce the accuracy loss in spatial transferability of a model.

To understand changes in cropland or crop types, it is necessary to have information from both past and current scenarios. Even though historical RS datasets for cropland mapping have been available (e.g., Landsat, MODIS), there is limited availability of field data to validate model predictions. At the same time, due to the missing field data for all spatial regions, the ability to transfer the model trained on one spatial region to another would help to close this gap. However, like in the case of temporal transferability, the evaluation of the model prediction is limited. For example, Kyere et al. (2019) and Orynbaikyzy et al. (2022) independently evaluated the temporal and spatial transferability of RS data models using RF algorithms. The authors of both studies attempted to enhance transferability by training on multiple year or location datasets. Regarding temporal domain transferability, Kyere et al. (2019) employed data from different years, while Orynbaikyzy et al. (2022) leveraged multiple location datasets (e.g., different states from Germany) for spatial domain transferability assessment. However, to our knowledge, no published studies evaluate all the possible crop mapping RS data model transferability scenarios – temporal, spatial, and spatial–temporal transferability, and their performance in terms of prediction quality.

Furthermore, the recently published study by Meyer and Pebesma, (2021) suggested a new approach to reporting the uncertainty of the ML model prediction in an unknown space. The study introduces the dissimilarity index (DI) computed by comparing the model's training data and data from the unknown space. In the second step, a threshold is set and applied to the DI values to mark where the model can successfully predict the value, and the so-called area of applicability (AOA) is defined. Applying the concept of DI and AOA can be beneficial to overcome the model's weakness and increase its robustness for transferring the model into new domains.

Instead of training models with multiple datasets (feeding all the possible variations of data), domain adaptation (DA) techniques could be another way to solve the model transferability problem in different domains. As Peng et al., (2022) summarised, multiple shallow and deep DA methods are available. Shallow DA methods, such as instance-based, feature-based, and classifier-based, are widely used due to their computational efficiency. However, their effectiveness in improving model transferability can vary depending on the specific application. On the other hand, deep DA methods utilise deep adversarial networks to align the distributions of source and target domains, potentially leading to better model transferability. A study by Peng et al. (2022) demonstrated the potential of DA methods to improve model transferability in two temporal RS datasets. However, it also highlighted the risk of exacerbating performance degradation if inappropriate DA methods are employed. Moreover, due to their computational complexity, the model training process can significantly slow down when DA methods are applied.

Before considering applications of the complex and computationally intensive DA methods, the changes in model performances during the transferability and the reasons for those changes need to be better understood. Therefore, this study defined two research questions, namely i) to what extent the crop type classification model performance is decreasing during model transferability and ii) what causes those performance changes when the model is trained with multiple Landsat datasets from the same domain (multiple years or locations). Based on the results of Kyere et al. (2019) and Orynbaikyzy et al. (2022), the authors of this study hypothesise that training the model with multiple Landsat datasets could enhance the model transferability between domains to map crop types in past periods. To answer the research questions and to prove the hypothesis, this study has been designed a) to

evaluate ML model transferability in three scenarios (temporal, spatial, spatial–temporal), b) to compare the effects of the training data and the ML algorithms for model transferability scenarios, and c) to understand possible reasons (e.g., precipitation, elevation) affecting the data, and limiting the model predictions.

2. Materials and methods

2.1. Study area

The study was conducted in two regions in Germany, namely North Hesse (NH) and Weser-Ems (WE) (Fig. 1). These regions were selected to consider different physical and geographic features like elevation, climate, terrain and diversity of landscapes. Moreover, both regions differ in their area. Additionally, they represent contrasting extents of the biogas sector development. NH experienced moderate biogas sector growth, while WE is a region with a substantial biogas expansion.

North Hesse region is located in the central part of Germany and consists of five districts – Kassel, Waldeck-Frankenberg, Schwalm-Eder, Hersfeld Rotenburg, and Werra-Meissner. The total area of the region is ca. 6900 km^2, and it is characterised by diverse landscapes and farming conditions (Kyere et al., 2019). The major soil region in North Hesse is mountains and hills predominantly of non-metamorphic sandstone, claystone and marlstone, and sedimentary rocks partly covered by loess. In the western part of the region mountains and hills predominantly of slates are found. Characteristic soils of the region are Cambisols, Podzols, Luvisols, Gleysols and Leptosols. The quality of soils measured in terms of Muencheberg Soil Quality Rating (Mueller et al., 2007) is predominantly very low and low (values up to 50) in western and eastern parts of the region in districts of Waldeck-Frankenberg, Hersfeld Rotenburg, and Werra-Meissner, while better quality soils with average (60–70) and high (70–85) values are found in the central part of the North Hesse region in Schwalm-Eder and Kassel districts. The favourable lands for farming are found mainly in flat valleys and plateaus with moderate slopes, which are often covered by loess of a substantial thickness (Wagner, 2011). The natural vegetation of North Hesse region is dominated by beech forests of low or moderately alkaline sites (Suck et al., 2014). Elevation ranges from 101 to 754 m with mean annual temperatures of 9–10 °C in the lowlands and 5–6 °C in the highlands. The mean annual rainfall ranges from 500 to 1300 mm (Kyere et al., 2019).

Weser-Ems region is a former government district of Lower Saxony in the northwest part of Germany. It consists of 12 districts and five district-free towns and has a total area of ca. 14,965 km^2. The region is characterized by high agricultural activity. In particular, the southern part of the region is dominated by pig and poultry farms, while the northern part hosts high shares of permanent grasslands serving the dairy cattle husbandry (Niedersächsisches Ministerium für Ernährung Landwirtschaft und Verbraucherschutz, 2022). Weser-Ems region is dominated by two soil regions: i) Holocene Coastal Plains with characteristic soils like Tidal Marsh and Regosols on the islands and along the in the North See coast in the northern part of the region and ii) older glacial drift areas with Cambisols, Podzols, Gleysols, Luvisols and Histosols in the rest of the region. High quality soils (Soil Quality Rating values of 70–85) are found in the coastal area in the districts of Aurich, Wittmund and Friesland, and in the western part of district Leer. The major part of agricultural soils in Weser-Ems region are of low (50–60) and average (60–70) quality. Extremely low (less than 35) and very low quality soils (35–50) are mostly found in districts Cloppenburg and Emsland (Mueller et al., 2007). Elevation ranges from 0 to ca. 95 m above sea level (Müller and Haberlandt, 2018). The coastal area of the Weser-Ems region is dominated by costal vegetation, in particular by sage meadows and salt marshes. Further main natural vegetation types in the northern part of the region are ash and sycamore wet forests and alluvial forests, mixed pedunculate oak forests, bog birch and black alder forests. In the southern part the natural vegetation is dominated by beech forests and mixed pedunculate oak forests of low-alkaline sites. In the south-west part of the region, predominantly in the district Emsland, the typical vegetation of highly acidic bogs is found (Suck et al., 2014). The area belongs mainly to the Northwest German lowlands climate model region with a mean average annual temperature of 8.6 °C and a mean average annual rainfall of 730 mm (Deutscher Wetterdienst, 2018).

2.2. Data

2.2.1. Reference data

Farmers consistently collect and submit the Integrated Administration and Control System (IACS) dataset as a component of the common agricultural policy subsidy payment program. Due to the data protection laws, the IACS data is not publicly available. However, upon the request of the relevant authorities, the data can be exceptionally made available to scientific institutions for research purposes. In the past, the IACS data has been successfully used for developing crop-type mapping tasks using satellite images as training and validation datasets (Blickensdörfer et al., 2022). Therefore, this study also employed the IACS data from the two study locations covering the period between 2010 and 2018. The selection of this period exclusively depended on the availability of the IACS data in the two study areas. The IACS data contained the GIS vector polygon layer, and each polygon had the following attributes: year and crop type that was grown in that polygon in the respective year. In this study, the crop types in the IACS data were grouped into five classes: maize, grasslands, winter crops, summer crops, and mixed crops. Additionally, this grouping was chosen to enable the use of the trained models for biogas development induced LULCC assessments beyond this study, e.g. by exploring the extent the increasing areas of maize have replaced other crops and grasslands, for example as reported by Levavasseur et al., (2023).

Even though the two selected regions are within the same country, they still showed distinct differences in dominating soils, types of natural vegetation, elevation and agricultural activities. The total selected crop field area for the WE was about 8800 km^2, and the entire crop field area in the NH was 2300 km^2. Similarly, the number of crop fields differed significantly between the two regions. The total number of crop fields in the WE was above 280000, and in the NH, it was above 98000. The distribution of the crop classes also showed discrepancies in the two regions (Fig. 2). The most prominent distinction was that about 25 % of crop fields in the WE region were maize, while in the NH region, it was less than 10 %. However, grassland and summer crops indicated similar proportions of fields in both regions. In contrast, about 35 % of fields in the NH region were winter crops, while only 15 % were winter crops in the WE region.

2.2.1.1. Satellite remote sensing data. The satellite image searching, downloading, and pre-processing were completed using the Google Earth Engine (Gorelick et al., 2017) Python application program interface and two Python libraries 'geemap' (Wu, 2020) and 'eemont' (Montero, 2021). Landsat scenes with less than 60 % cloud coverage (from Landsat 5, 7, and 8 satellites) which covered both study regions, were retrieved for each year (between 2010 and 2018) from March to October. Even though there are slight differences in spectral ranges of the three Landsat sensors, it was confirmed that there is no significant impact on the classification results due to these spectral differences (Flood, 2014). The Landsat 5 data was available until May 2012, and Landsat 8 data was available from 2013. The Landsat 7 data has been available since 1999, but due to its scan line corrector failure, all the images since 2003 contained gaps (Wulder et al., 2019). In this study, the data from Landsat 7 data was not corrected to remove the gaps, and those gaps were considered as no data pixels similar to cloud and shadow areas. The retrieved Landsat scenes were Collection-2 Level-2 products that provided atmospherically corrected surface reflectance

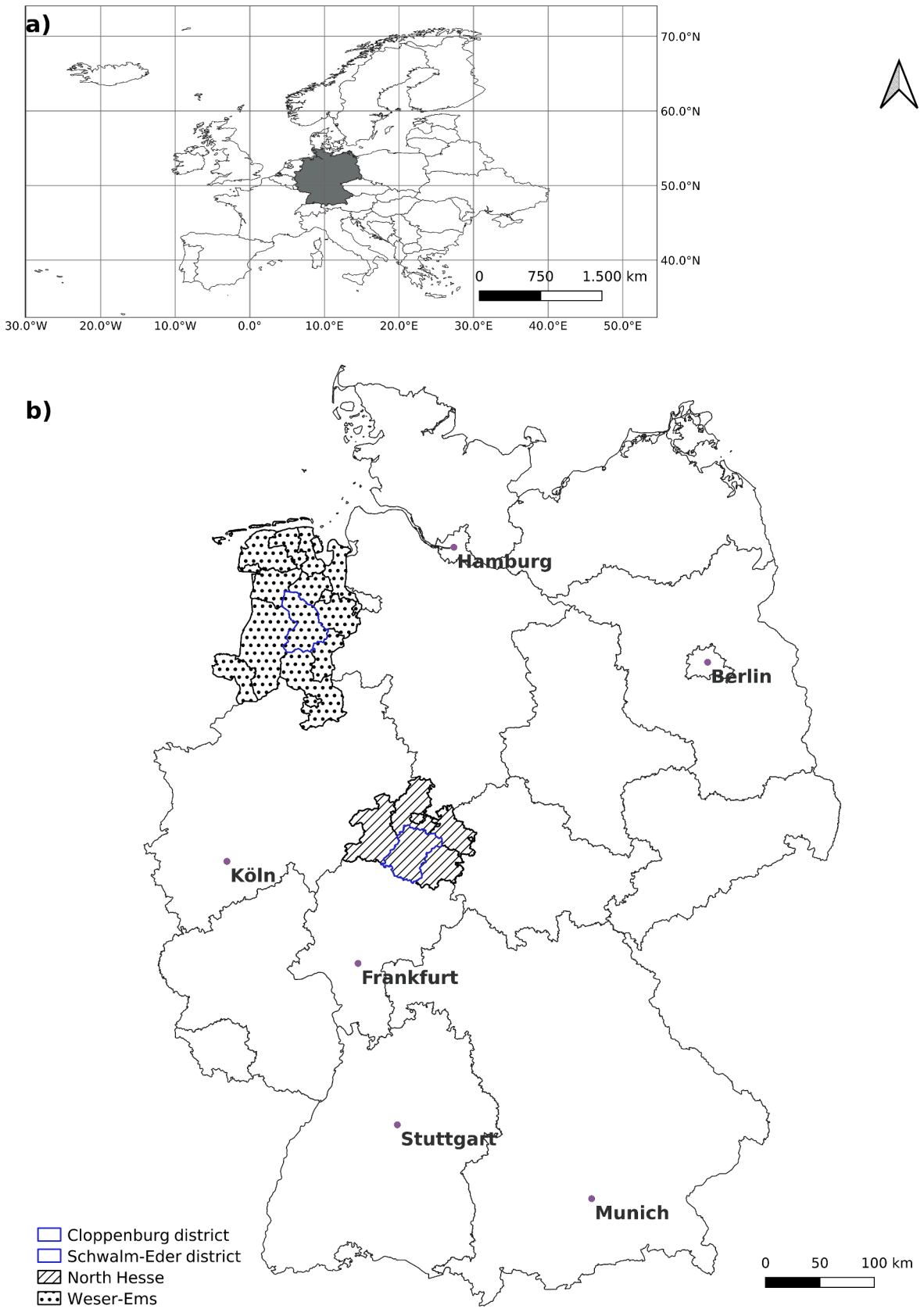

Fig. 1. Location of Germany (grey-coloured country) in the continent Europe (a) and location of two study areas (pattern filled area) within Germany, including locations of major cities in Germany (b). The blue-outlined areas were used as testing areas for the reference scenarios in each respective modelling exercise. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

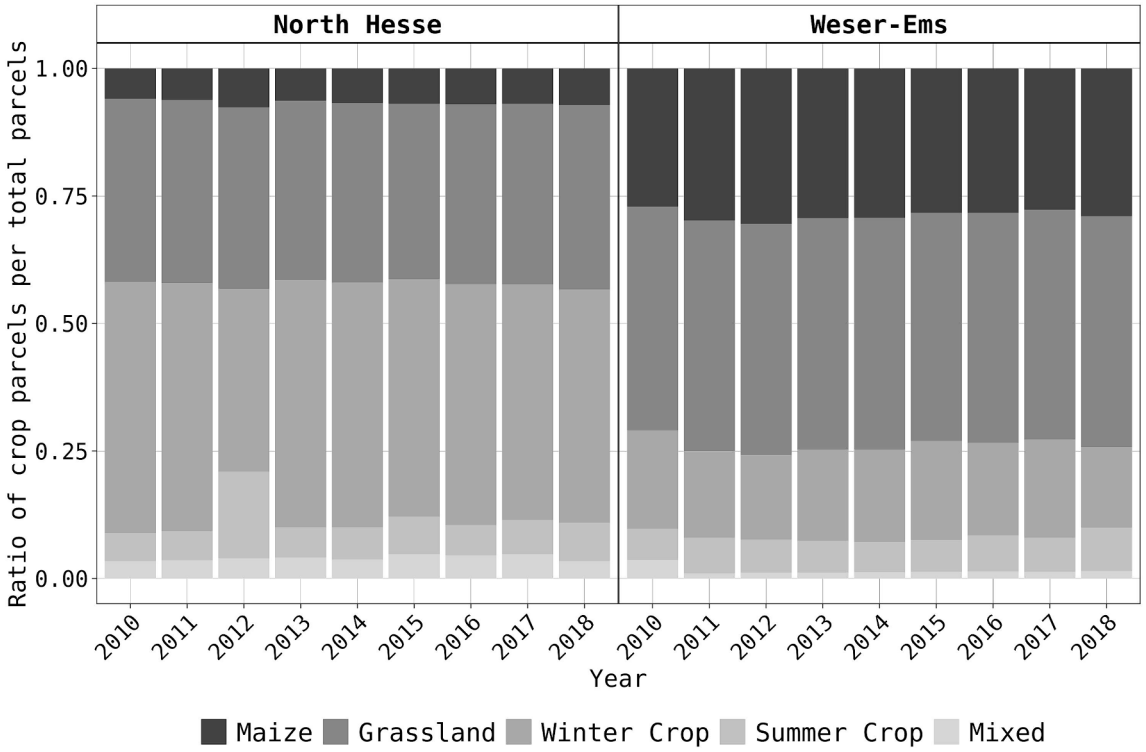

Fig. 2. Crop parcel ratio distribution in the two study regions between years 2010 and 2018.

data, including the cloud mask layer (Landsat Level-2 Surface Reflectance Science Product courtesy of the U.S. Geological Survey) (Crawford et al., 2023).

First, these image scenes were masked using the cloud masking algorithm from 'eemont' library and then scaled the images (between 0 and 1) using offset and gain values from metadata. Since Landsat 5 sensor (thematic mapper – TM) only contained six spectral bands (B1: blue, B2: green, B3: red, B4: near- infrared1, B5: near-infrared1, and B7: mid- infrared) the same spectral bands were selected from the Landsat 7 (enhanced thematic mapper plus – ETM+) and Landsat 8 (operational land imager – OLI). In addition to the spectral bands, four spectral vegetation indices (VIs) were computed following Kyere et al. (2019) (normalised vegetation index – NDVI, enhanced vegetation index – EVI, optimised soil-adjusted vegetation index – OSAVI, and normalised difference moisture index – NDMI). At the end of this step, each image scene contained six spectral bands and four VIs. Then all the annual image scenes were categorized into each bimonthly period (i.e., March-April, May-June, July-August, September-October), and the median image per each two-month (temporal aggregation) period was computed. It resulted in four bi-monthly median images per year, each containing ten image layers (6 bands + 4 VIs). Finally, those four median images with ten layers were downloaded for each year.

According to the rule of thumb and minimum mapping unit, at least nine contiguous pixels are required to identify an object. To avoid mixed pixel problems, only crop fields bigger than 0.9 ha (30 m x 30 m x 10 = 9000 m^2) were selected in this study. Satellite data of each year were overlaid with corresponding IACS data, and mean values of each band and VI per each crop field polygon were extracted. Data preparation and extraction were done using R programming. Each polygon contained 40 mean values (10 layers x 4 time points).

2.3. Model development for the reference case

This study employed two machine/deep learning (ML/DL) algorithms for evaluating model transferability scenarios. The RF ML algorithm was one of the algorithms explored in this study because it has already been widely applied by the previously mentioned studies (Kyere et al., 2020; Orynbaikyzy et al., 2022) that examined spatial and temporal transferability of crop type mapping. The RF algorithm is an ensemble tree-based algorithm (Breiman, 2001) which needs input as a 1D array. In this study, the input vector for the RF algorithm was a 1D array with 40 elements (Fig. 3).

The convolution neural network (CNN) was the second algorithm examined in this study. CNN is a state-of-the-art deep-learning modelling method that showed the potential for land cover and crop type classification using satellite time series data (Pelletier et al., 2019). Time series data are usually 1D data, and application 1D CNN is the most common application in crop-type classification using satellite time series data (Pelletier et al., 2019). So, this study applied a spectral-temporal aggregation that converted 1D time series data into 2D array data to apply 2D CNN, which was a step forward from the method suggested by Pelletier et al., (2019). Initially, one instance of input data in this study was the 1D array with 40 elements. So, it was converted to a 2D array with ten rows and four columns, where each row represents a band or VI, and each column represents the temporal value of the corresponding band or VI for four temporal points per year (Fig. 4). This new 2D array served as input into the 2D CNN, where the convolution filters can simultaneously identify the combination of spectral and temporal

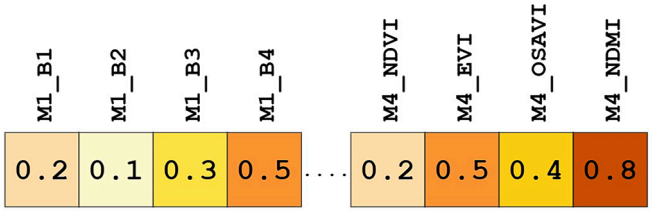

Fig. 3. Time series data as a 1D array (Not all 40 values are represented). M stands for month, and the number next to the letter M is for the month number. B stands for Landsat image band number. NDVI: Normalised difference vegetation index, EVI: Enhanced vegetation index, OSAVI: Optimised soil adjusted vegetation index, NDMI: Normalised difference moisture index.

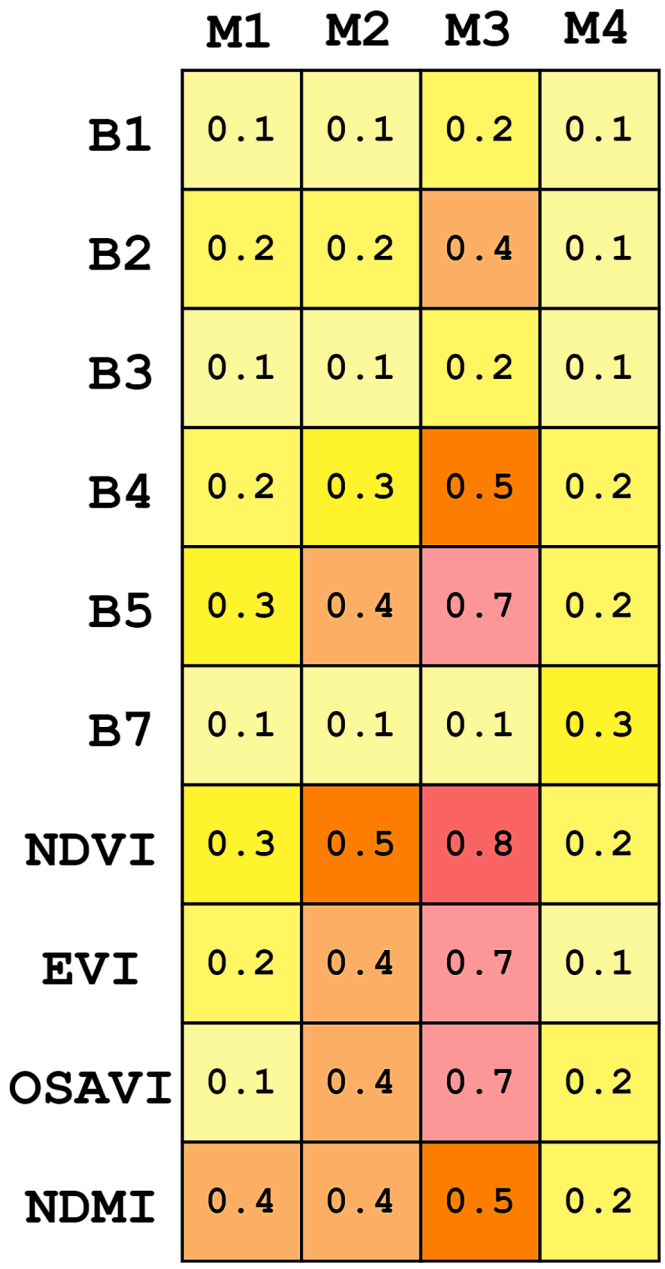

Fig. 4. Time series data as a 2D array (All 40 values are included). M stands for month, and the number next to the letter M represents the number of the month. B stands for Landsat image band number. NDVI: Normalised difference vegetation index, EVI: Enhanced vegetation index, OSAVI: Optimised soil adjusted vegetation index, NDMI: Normalised difference moisture index.

patterns. This study exemplarily focuses on mentioned two algorithms, which have already shown the potential of crop-type mapping applications with time series RS data, representing both traditional ML and DL aspects.

The model development and evaluation were conducted in Python using 'sklearn' and 'tensorflow' libraries. Since this study contained two classification algorithms (RF, CNN), and two study sub-regions (NH, WE), a total of four models were trained. The summary of the trained models is given in Table 1. The hyperparameters of the RF models were found using a subsample data set. According to that, RF hyperparameters were *max_features* = *'sqrt'*, *criterion* = *'entropy'*, *min_samples_leaf* = *1*, *min_samples_split* = *8*, and, *n_estimators* = *600*. The model training was done with parallel processing.

The 2D CNN model consisted of two convolution blocks followed by a fully connected flatteneden layer. In each convolution block double convolution with the 'ReLu' activation function was applied and batch normalization was applied in between. The convolution filters always were 3 × 3 kernel with 'same' padding and one stride value. At the end of each convolution block 2D max pooling was done. The fully connected layers contained 25 % dropout. For both class levels, above 96,000 trainable parameters were in the 2D CNN models. The model optimizer was set to stochastic gradient descent (SGD) with learning rate of 0.0001 and momentum of 0.9. The categorical cross-entropy loss was employed as the loss function in the model training. The model training was done for 50 epochs using GPU processors.

Table 1
Overview of trained models.

Spatial region & period of the training data	Machine learning algorithm	Model name
North Hesse 2013 – 2018	Random forest	NH-RF
	2D Convolution Neural Network	NH-CNN
Weser-Ems 2013 – 2018	Random forest	WE-RF
	2D Convolution Neural Network	WE-CNN

In the model training process, spatial–temporal subsets from both training data were held out for model testing purposes. The respective held-out datasets were Schwalm-Eder district (DE735) data in 2017 from the NH and Cloppenburg district (DE948) data in 2017 from the WEs. The held-out dataset was applied to the trained model. The model predicted new labels, and the predicted values were compared against the actual values. The following model evaluation metrics were computed: overall accuracy (OA) (Eq. (1)), class-level F1 (Eq. (2)), and Macro F1 (mF1) (Eq. (3) values. Since this was the usual model training and testing procedure without transferability scenarios, it was considered a reference case scenario. In this case, the evaluation metric values are the reference OA, reference class level F1 and reference mF1.

$$Overallaccuracy(OA) = \frac{No.ofcorrectlyclassifiedsamples}{Totalsamples} \times 100 \quad (1)$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

$$MacroF1(mF1) = \frac{\sum_{i=1}^{i=n} F1_i}{n} \quad (3)$$

Class-level F1 in multiclass problem was computed similarly to binary classification problem by considering one class vs rest classes approach. In Eq. (2)) TP is truly positive, FP is false positive, and FN is false negative. mF1 (Eq. (3)) is the arithmetic mean of each class-level F1 score.

2.4. Model testing for transferability cases

The trained models were tested in three model transferability scenarios to evaluate the temporal, spatial, and spatial–temporal transferability of the models (Fig. 5). The summary of the dataset tested in each transferability scenario is explained in Table 2. First, the data from the same region but in a separate temporal period was evaluated in the temporal transferability case. Next, the data from the same period but in other spatial regions was assessed in the spatial transferability case. Finally, data from distinct spatial regions and temporal periods were tested in the third spatial–temporal transferability assessment.

In each transferability case, model predictions were compared with the actual labels, and the aforementioned evaluation metrics (i.e., OA, mF1, class-level F1) were computed. Furthermore, to assess changes in each transferability case against the reference case, a percentage of change of mF1 and class-level F1 values was computed (Equation (4).

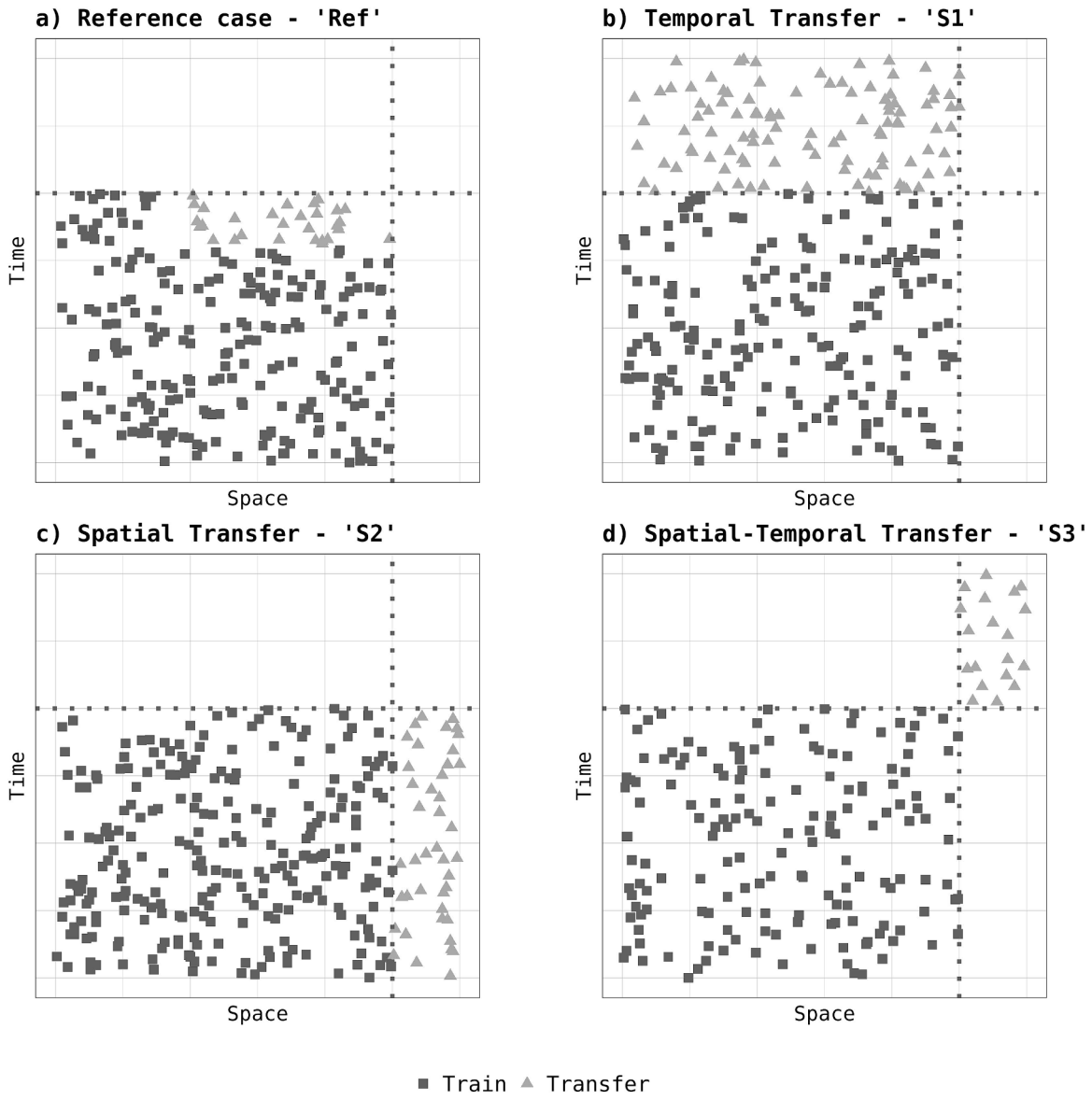

Fig. 5. Schematic representation of the a) reference case (no transfer outside the spatial and temporal domain), b) temporal transfer (S1), c) spatial transfer (S2), and d) spatial–temporal transfer (S3). The dotted line represents the training data boundary of each domain. The black squares represent the distribution of the data used for model training, and the grey triangles indicate the distribution of test data where the model is transferred to new data. In each transferability scenario, the test data (transfer) would always be outside the training data boundary of the time and/or space domain.

Table 2
Assessed scenarios and respective training and testing datasets.

Scenario	Training dataset	Testing dataset
Reference (Ref)	NH data between 2013 and 2018	Subset data from training dataset from Schwalm-Eder district in 2017
	WE data between 2013 and 2018	Subset data from training dataset from Cloppenburg district in 2017
Temporal transferability (S1)	NH data between 2013 and 2018	NH data between 2010 and 2012
	WE data between 2013 and 2018	WE data between 2010 and 2012
Spatial transferability (S2)	NH data between 2013 and 2018	WE data between 2013 and 2018
	WE data between 2013 and 2018	NH data between 2013 and 2018
Spatial-temporal transferability (S3)	NH data between 2013 and 2018	WE data between 2010 and 2012
	WE data between 2013 and 2018	NH data between 2010 and 2012

$$Metric_{Change} = \left(\frac{Metric_{Reference} - Metric_{Transferability}}{Metric_{Reference}}\right) \times 100 \quad (4)$$

2.5. Model transferability exploration

To explore the potential reasons for changes in model evaluation metrics in the transferability cases, first, the DI concept from Meyer and Pebesma, (2021) was applied. The DI values can explain the dissimilarity of the predictor variables compared to the training data. The DI threshold value was defined based on the computed DI values for the training dataset. The DI values of the test datasets were separated using the threshold value, where data below the threshold is considered true for AOA and the data above the threshold value will be regarded as false for AOA. The DI and AOA calculation was done using the 'CAST' package in R (Meyer et al., 2023). In each transferability scenario, the percentage of data that came under true for AOA was linked with the model accuracies, and their relationship was assessed. According to the relationship, possible reasons for variation in model performances were

explained.

Moreover, to explain the changes in model performance, other possible factors (i.e., parcel size, elevation/slope, rainfall/temperature) in each transferability case were compared against the reference case. Above mentioned factors can have significant impact on model accuracy. For example, Kyere et al., (2019) highlighted that smaller size fields tend to result in lower accuracy in crop type mapping compared to bigger fields due to high probability of mixed pixels. Similarly, Orynbaikyzy et al., (2022) pointed out that factors such as elevation and soil quality could also affect model's prediction on the new location data. Further, crop type mapping analysis by Blickensdörfer et al., (2022) also showed that climate and season variable could impact the crop type mapping using RS time series data. The field size data was computed from IACS data. Elevation data was acquired from the 30 m global Shuttle Radar Topography Mission (SRTM) digital elevation model (Shuttle Radar Topography Mission Global, 2013), and the slope was computed. The weather data (temperature, precipitation) was retrieved from the 'global surface summary of the day' database (National Climatic Data Center, 2021). Graphical and statistical comparisons using boxplots were made to understand the difference between reference and transferability case data.

3. Results

3.1. Landsat time series data

A selected VI (NDVI) bi-monthly average time series for each crop class is shown in Fig. 6 for the two regions and two distinct years. Mostly, there was a distinct time series pattern among the five crop classes explored in this study. The maize class had lower VI values between March and June, then showed maximum VI values during July-August. A similar pattern could be seen for the summer crops in the WE. However, the summer crop VI time series for the NH showed maximum values in both May-June and July-August. Furthermore, a variation between years can be observed too. For the winter crops, the highest VI values were in May-June, and, after that, they showed reduced values in other months. Additionally, grassland showed slight VI value changes during the whole period. Nevertheless, the grassland fields in the NH indicated a higher fluctuation of VI value compared to the WE's grassland fields. The VI time series for mixed crop class in the NH showed a comparable pattern to winter crops, and mixed crops in the WE showed an equivalent pattern to summer crops.

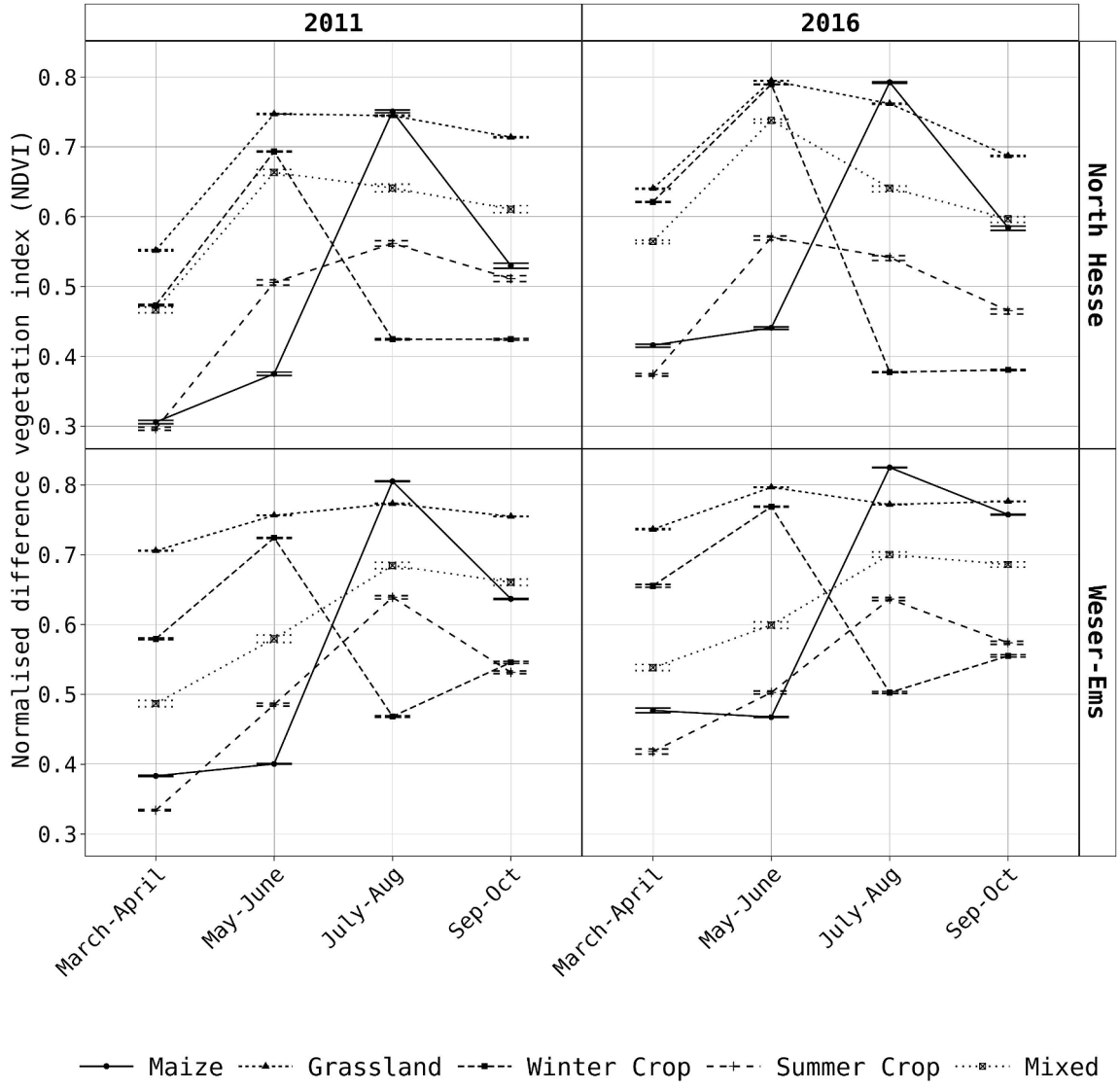

Fig. 6. Average normalised difference vegetation index (NDVI) bi-monthly time series for each crop class in two different regions and two selected years. Lines in each point indicate the upper and lower confidence intervals of each average NDVI value.

3.2. Model performances in the reference case (Ref)

All models showed more than 88 % OA regardless of utilised training data (i.e., NH data, WE data) or ML algorithm (i.e., RF, CNN). Table 3 summarises the OA and mF1 values of each model in the reference case. NH data models showed similar mF1 values for both algorithms (0.76). However, for the WE data model with CNN showed a slightly higher mF1 value than RF (CNN = 0.74, RF = 0.72).

From all four models, the winter crop class obtained the highest class-level F1 values (over 0.95) (Fig. 7). Maize and grassland classes also showed above 0.85 class level F1 values. However, the summer crop class showed above 0.75 F1 values for models trained with NH data, while the same class's F1 value from the models trained with WE data was 0.65. The mixed crop class demonstrated the lowest class level F1 values in all models. Nevertheless, there was no significant difference in F1 values between the two ML algorithms.

3.3. Model performance for the transferability scenarios

The models trained with WE data showed an apparent decrease in performance from S1 to S3 (Table 4). For example, the OA of the three scenarios of the RF model trained with WE data were 85 %, 82 %, and 70 % for the S1, S2, and S3 scenarios, respectively. However, the models trained with NH data using the RF algorithm demonstrated slightly different results than those with CNN, where the S1 transferability scenario obtained the lowest performance. In the S1 scenario, both ML algorithms showed similar mF1 values for the NH data models, while narrowly better performance was obtained by CNN for WE data models. Nevertheless, RF models obtained scarcely better mF1 values than CNN models in S2 and S3 scenarios in both NH and WE data models.

3.4. Comparison of reference and transferability cases

The model performance's metric values from each transferability scenario were compared against those of the reference case, and the percentage change of the metric values was computed. The percentage difference of the mF1 values was plotted in Fig. 8. The comparison distinctly revealed that for all models, the model performance is decreasing when the model is transferred to different complex domains. As mentioned earlier, the performance of the model trained with WE data decreased from the reference case to S3 scenarios in an orderly manner (Ref > S1 > S2 > S3). For the S1 scenario, the mF1 percentage difference was below −6 % for both ML algorithms. The RF models trained with WE data resulted in lower performance losses during S1 and S2 scenarios than CNN models. In the S3 scenario, the models trained with WE data at CL2 level showed −19 % and −24 % of mF1 differences for the two ML algorithms.

The RF models trained with NH data demonstrated a unique pattern where the mF1 difference was similar in all three scenarios (−15.5 % on average). In contrast, CNN models for the same setup displayed decreasing model performance from the reference case to the S3 scenario (Ref > S1 > S2 > S3), which was similar to the models trained with WE data.

Class level F1 values presented different crop-specific patterns for the three different model transferability scenarios (Fig. 9). The maize class showed a positive or no difference in F1 values in three transferability scenarios for the model trained with NH data. However, the model trained with WE data indicated decreasing F1 values from S1 to S3 scenarios. Comparable to the maize class, the grassland class also showed a similar F1 value pattern in the two model types. In contrast, summer and winter crop classes F1 values from the model trained with NH data demonstrated decreasing patterns from S1, S2, and S3. Nevertheless, the model trained with WE data did not show such a prominent pattern for those two classes.

Table 3
Summary of the model performance in the reference case. NH: North Hesse, WE: Weser-Ems, RF: random forest, CNN: 2D convolution neural network.

Model name	Overall accuracy (%)	Macro F1
NH-RF	90.2	0.765
NH-CNN	89.7	0.758
WE-RF	88.3	0.722
WE-CNN	90.1	0.740

3.5. €Model transferability and area of applicability

The DI-based AOA values were computed for each observation in the reference case and three transferability scenarios. According to the percentage of the observations where AOA was True was linked against the OA values (Fig. 10). The models trained with NH data indicated that 88 %, 90 %, 89 %, and 78 % of the tested data were within the AOA for Ref, S1, S2, and S3 scenarios. However, those values did not show a clear pattern with the complexity of the transferability scenarios. Nevertheless, 93 %, 90 %, 90 %, and 87 % of observations from Ref, S1, S2, and S3 scenarios were considered within the AOA for the models trained with WE data. It explained that the amount of data points that were similar to the data that used to train the model were decreased when the complexity of transferability cases increased. The same pattern was shown in the OA values from the model trained with WE data (Table 4). Therefore, the correlation coefficient between OA and the percentage of True AOA observations was significantly positive (0.9).

3.6. (Potential) influences of spatial and temporal conditions settings

Crop parcel size, elevation, and slope differed among spatial regions, which may be a potential cause for performance reduction in S2 (spatial) and S3 (spatial–temporal) transferability scenarios. As shown in Fig. 11, the field sizes of all crop classes were bigger in the WE region compared to the NH region. In S2 and S3 scenarios, the models trained with the NH data, which contained smaller fields, obtained better model performances when transferring the model to a bigger field size region.

Since the WE region is located in a flat area in northern Germany, the mean elevation of most of the crop fields in the WE was between 0 m and 50 m (Fig. 11). Similarly, the slope percentage in the crop fields was also near zero. In contrast, the crop fields in the NH region had a mean elevation of about 300 m, and the percentage slope of the fields was 10 – 15 %. When the model trained with NH data (NH-CNN) was tested with data from the WE region (under the S2 scenario), the model performance decreased by 20 %. However, when the model trained with WE data was transferred to the NH region, less than 15 % of performance loss occurred. The possible explanation is that models trained with relatively flat area data could perform well also in highly undulated landscapes.

In the S1 and S3 scenarios, temporal periods were dissimilar between the data used for model training and testing in the transfer region. Weather patterns (e.g., temperature, precipitation) and crop calendars are possible factors that may vary between two different periods. The mean air temperature and precipitation patterns of the two regions and two different periods are illustrated in Fig. 12. However, there was no apparent difference or shifts in temperature and precipitation patterns either between regions, nor between references (2013 – 2018) or transfer years (2010 – 2012). Nevertheless, when NH models were transferred to a new temporal domain, their performances decreased by about 15 %. In comparison, WE data models showed only a 5–6 % reduction in model performance. Further, the crop calendar (i.e., sowing date and harvesting date) within the same region could differ due to weather conditions in distinct years, which might influence the model predictions by altering standard phenological patterns.

4. Discussion

Ascribable spatial and temporal changes across the data can reduce

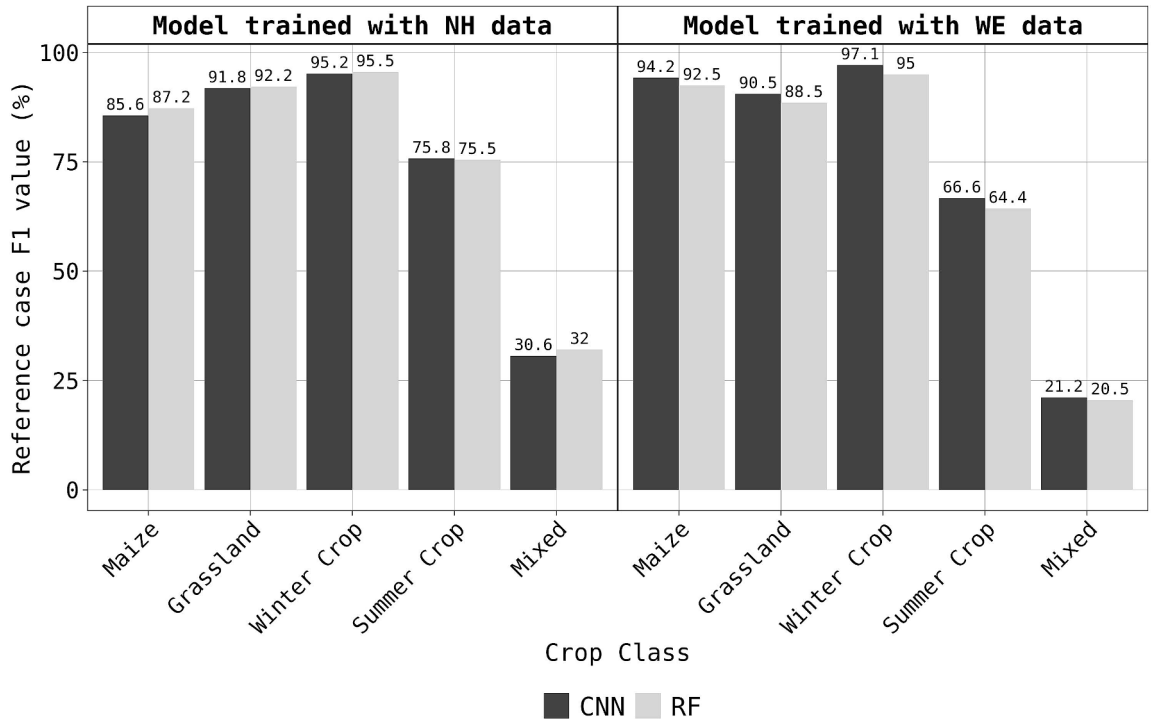

Fig. 7. The class-level F1 values bar graphs for the reference case. CNN: 2D convolution network algorithm model, RF: random forest algorithm model, NH data: North Hesse data, WE data: Weser-Ems data.

Table 4
Model performance summary for the temporal (S1), spatial (S2), and spatial–temporal (S3) transferability scenarios. NH: North Hesse, WE: Weser-Ems, RF: random forest, CNN: 2D convolution neural network.

Model name	Overall accuracy (%)			Macro F1		
	S1	S2	S3	S1	S2	S3
NH-RF	74.1	85.0	84.4	0.645	0.645	0.647
NH-CNN	83.7	82.8	80.5	0.678	0.614	0.598
WE-RF	85.4	82.2	70.0	0.698	0.678	0.588
WE-CNN	83.6	75.5	62.1	0.695	0.643	0.563

the model performances when the models are transferred to unfamiliar spatial and temporal domains. The unavailability of adequate data that could explain all the possible variabilities in spatial and temporal domains always makes it challenging to build a generalised model that works across domains. Consequently, models trained on one domain are often transferred to a new domain despite a performance loss. However, understanding the extent of performance reduction during the model transferability is essential for analysing the predicted values. Therefore, this study was designed to quantify or assess how much the model performance changes when the trained model is transferred to three domains (S1, S2, and S3) for crop type mapping application using Landsat time series data. The findings from this study demonstrated that model performances decreased in all three transferability scenarios compared to the reference case (no-transfer case), and the order model performance reduction were temporal, spatial and spatial–temporal scenarios in ascending order of losses.

Temporal transferability was the first scenario that assessed the model transferability to predict crop classes in the same spatial domain as the model trained but during another period. Usually, crop management practices might be similar in the same region, and the distribution of the crop types within the regions could be also similar. However, the crop calendar (i.e., sowing date and harvesting date) could differ even within the same region due to weather conditions in distinct years. Multiyear data could usually be employed to train the model to account for these variations, and this study also followed a similar approach by training models using data from five years (2013 – 2018). Even though the models learned from the 5 years data, the model performances were reduced by up to 15 %. These results align with the results reported by Kyere et. al., (2019), where trained crop-type classification models with multiyear Landsat data also showed reduced performance for prediction on new year data. In comparison to the crop-type mapping, Praveen et al., (2019) reported that the overall accuracy of LULC mapping (e.g., built-up, cropland, fallow land, grassland, water) was reduced by only about 1 % compared to the reference case.

However, the results showed that two models from two regions indicated completely different accuracy losses. The temporal transferability models with NH data showed lower performances than the models with WE data because the field size in the NH was comparatively small. Kyere et. al., (2019) also explained that a smaller field size lowers the model's prediction accuracy. The possible reason for that could be that RS image mean values extracted from smaller fields were only based on a small number of pixels, and those pixels could also be affected by boundary effect that could have the mixed pixels instead of pure crop pixels.

Based on Fig. 12, visual changes in climate variables were not found, and the authors did not explore this in-depth in this study. However, further research to investigate this matter and to find out the link between model transferability and changes in climate variables or how to include them in the model is needed. The climate shift in two different regions or periods could also impact the different phenological stages of the crops, which could limit the model transferability. Further, the changes in cultivars, farming practices, such as the preparation of land, and post-harvesting practices due to different policy implementations and technological evolvement in various years could also be reflected in the vegetation phenology (Ghazaryan et al., 2018). These changes in phenology pose a challenge in achieving consistent RS time series patterns across different years (Blickensdörfer et al., 2022). Therefore, more research is needed regarding the consideration of environmental, climatic and geographic variables in RS data analysis concerning the model performance in transferability scenarios (Blickensdörfer et al.,

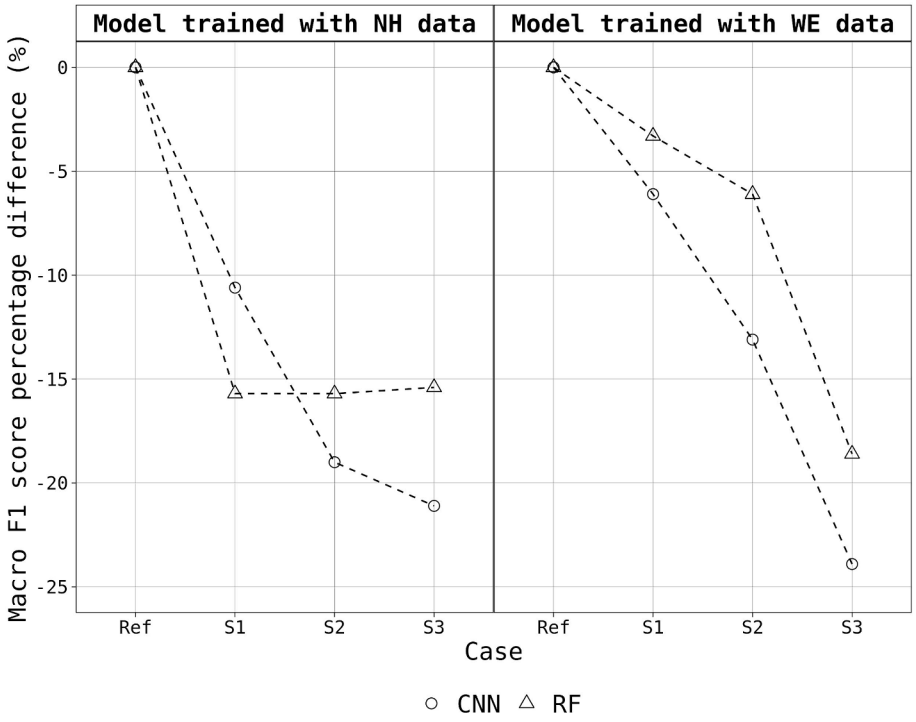

Fig. 8. Macro F1 (mF1) score difference as a percentage for three transferability scenarios. S1: temporal transferability, S2: spatial transferability, S3: Spatial-temporal transferability, NH: North Hesse, WE: Weser-Ems, CNN: 2D Convolution neural network, RF: Random forest.

2022). For example, applying growing degree days into classification models could tackle shifts in phenology due to climate differences in two spatial regions or temporal periods (Nyborg et al., 2022a).

Apart from the environmental reasons, applying a bi-monthly Landsat time series (four observations per year) could also limit the temporal transferability where changes in crop-specific temporal patterns would not be prominent to distinguish the analysed crop types within source and target temporal domains. The temporal transferability of RF models with four observations (bi-monthly) showed an average of 79.8 % accuracy, while RF models from Kyere et al., (2019) with two observations showed 77.3 % average accuracy. This explained that the densification of the RS time series could help to distinguish crop types easily and improve temporal transferability. Supporting the same argument, Blickensdörfer et al., (2022) reported that densified RS time series using Sentinel 2 and Landsat data improved crop type classification. However, to go back in time to map crop types to analyse the effect of biogas development on agricultural land cover during the last 20 years, Landsat stands as the sole viable choice (Kyere, 2020).

The model trained with data from one spatial region was tested on data from another under the second transferability scenario (S2). This study showed that when the model trained with data from multiple districts from similar regions was transferred to a new spatial region, the absolute Macro F1 values were changed between −0.04 and −0.14. In comparison, the crop type (11 classes) classification models with Sentinel-2 single crop year data also confirmed that the spatial transferability of the models could decrease the macro F1 value between 0.02 and 0.15 (Orynbaikyzy et al., 2022). The performance loss for spatial transferability can be likely attributed to changes in environmental, climate, and management conditions. For example, Orynbaikyzy et al., (2022) reported that the crop classification model transferability within different spatial locations inside Germany could be challenging because farmers selected their cultivars according to the environmental conditions. For example, different cultivars from the same crop type are cultivated in distinct regions due to contrasts in soil quality, which may have dissimilar phenological patterns that could increase the complexity. Additionally, the distribution of crop types in two regions could also reduce model performances. For example, maize was the most prominent crop in the WE region and when the model trained with WE data was transferred to the NH region, the less prominent maize share showed substantial performance loss in the maize class. In comparison, similar performance loss was observed for the winter crop class when the model was trained with NH data, which contained a higher share of winter crops transferred to the WE region, which had a lower share of winter crops. As suggested by Rusňák et al., (2023), this problem could be overcome by consideration of crop type distribution in the models. Further, incorporating data from larger regions might decrease the crop type distribution problem and increase the model transferability in the spatial domain (Johnson and Mueller, 2021). However, having label data over larger regions is mostly the bottleneck in crop type mapping exercises.

In the third transferability scenario (S3), the complex shifts in both temporal and spatial domains were assessed. The models predicted values in an unexplored spatial region and distinct period and it showed that the model performances decreased by about 25 %. These results confirmed that it is very challenging to transfer model across both spatial and temporal domains. DI and AOA-based calculation results followed a similar trend. The implementation of DI-based AOA demonstrated the potential explanation for the uncertainty in the model transferability. A positive relationship was noticed between the quantity of observation under AOA and the model accuracies in the transferability scenarios (Fig. 10). Consequently, when label data is unavailable to assess the model transferability performance, AOA may offer a means of comprehending the prediction uncertainty in the new domain (spatial and/or temporal) (Meyer and Pebesma, 2021).

As mentioned earlier, this study employed the Landsat time series data from Landsat 5, 7 and 8 satellites to map events in past periods. The temporal target domain data (between 2010 and 2012) were only from Landsat 5 and 7, and the source domain data (between 2013 and 2018) contained data from Landsat 7 and 8. Even though Flood, (2014) reported that there was no impact of multi-sensor Landsat satellite bands for classification purposes, there could be an impact from changes in the satellite sensors to the model transferability. On the other hand,

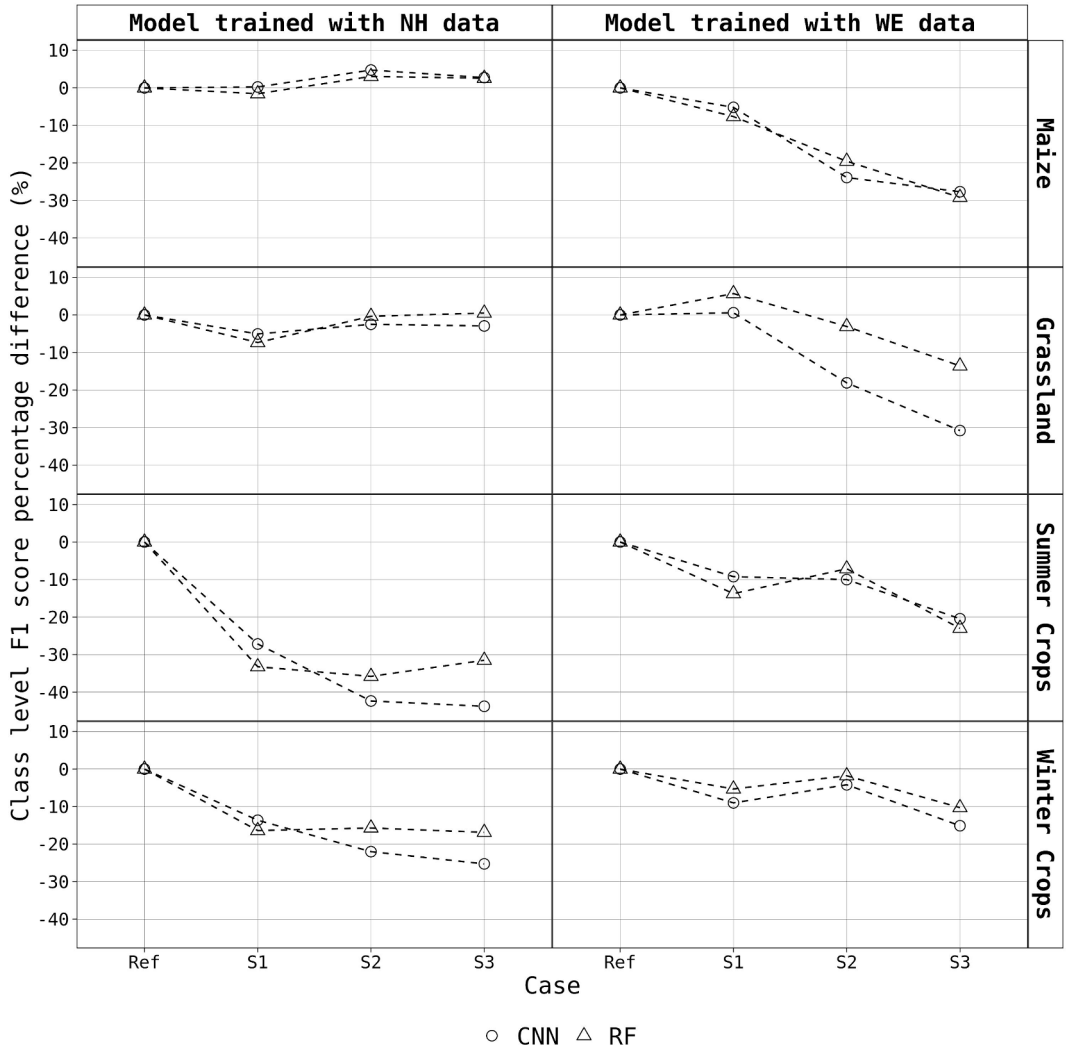

Fig. 9. Class level F1 score difference as a percentage for three transferability scenarios. S1: temporal transferability, S2: spatial transferability, S3: Spatial-temporal transferability, NH: North Hesse, WE: Weser-Ems, CNN: 2D Convolution neural network, RF: Random forest.

supporting that argument, Frantz et al., (2023) reported that spatial and temporal metrics derived from multiple Landsat sensors tend to contain uncertainties between years, and the quality of the observation increased from the past to the present when advanced satellite sensors were introduced. Furthermore, the computation of the median image from each two-month observation could have affected the data because the number of scenes per area in a given two-month period could differ due to the availability of scenes with clear sky observations. However, according to Frantz et al., (2023) the number of scenes with clear sky observation is not more important than seasonal data distribution, which this study tried to keep consistently.

RF was one of the ML models employed in this study where derived spatial–temporal RS values were directly input as predictor variables followed. The authors did not explore other ML models since the RF was already proven to be one of the best ML models for crop type mapping tasks with temporal RS variables (Kyere et al., 2019; Orynbaikyzy et al., 2022). As ML models tried to find the nonlinear relationships of the data in the hyperspace, it was not easy to adjust the models to adapt to the source and target domain. However, applying shallow DA techniques might be helpful to map data from both source and target domains to the shared domain before applying it to ML models (Peng et al., 2022). However, this was not a part of this study, and could be explored in the future research work. In addition, evaluating other available ML models (e.g., XGBoost) would also be recommended to find out how DA methods will deal with different ML models when transferred to different spatial and temporal domains.

In terms of DL, this study explored the spectral-temporal guidance CNN model, which showed better results in the no-transfer scenario than the RF model. Similar results were reported by Pelletier et al., (2019), who used similar CNN models and obtained better OA values than respective RF models. However, limited temporal observation in the dataset (four per variable) might be restricted the potential of CNN models, where more temporal observation showed better results (Pelletier et al., 2019). Furthermore, the developed CNN models showed poor transferability performances compared to RF, probably because multiple temporal observations were required for better understanding the difference between source and target domains. On the other hand, this study did not explore the application of deep DA methods with CNN models and as mentioned earlier, it will be the next step of our research. Therefore, using novel deep transfer learning could increase the model transferability using CNN and other DL models (Ma et al., 2024; Rußwurm and Körner, 2020). One example would be, self-supervised pretraining of the data using novel DL methods to solve the problems like temporal shift, spectral noise, inconsistencies in the time series data (Nyborg et al., 2022b; Xu et al., 2024; Yuan and Lin, 2021). Even though there has been massive progress in this field in recent years, there are still many challenges. For example, increased computational costs, the limited availability of adequate training datasets and incorporation with

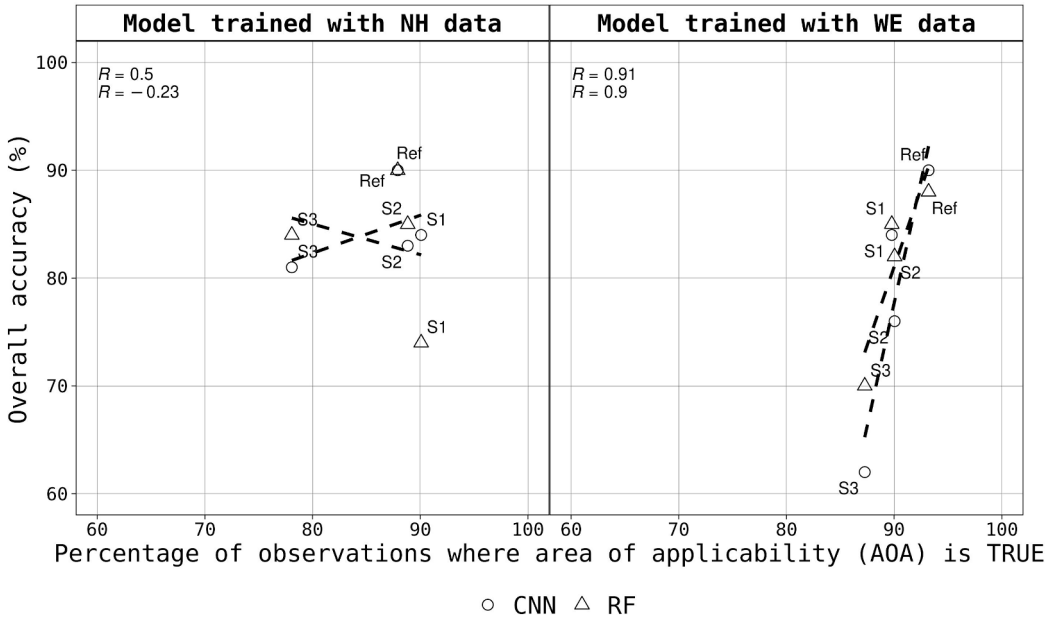

Fig. 10. Scatter plots between the percentage of observation under the area of applicability true and overall accuracy. S1: temporal transferability, S2: spatial transferability, S3: Spatial-temporal transferability, NH: North Hesse, WE: Weser-Ems, CNN: 2D Convolution neural network, RF: Random forest.

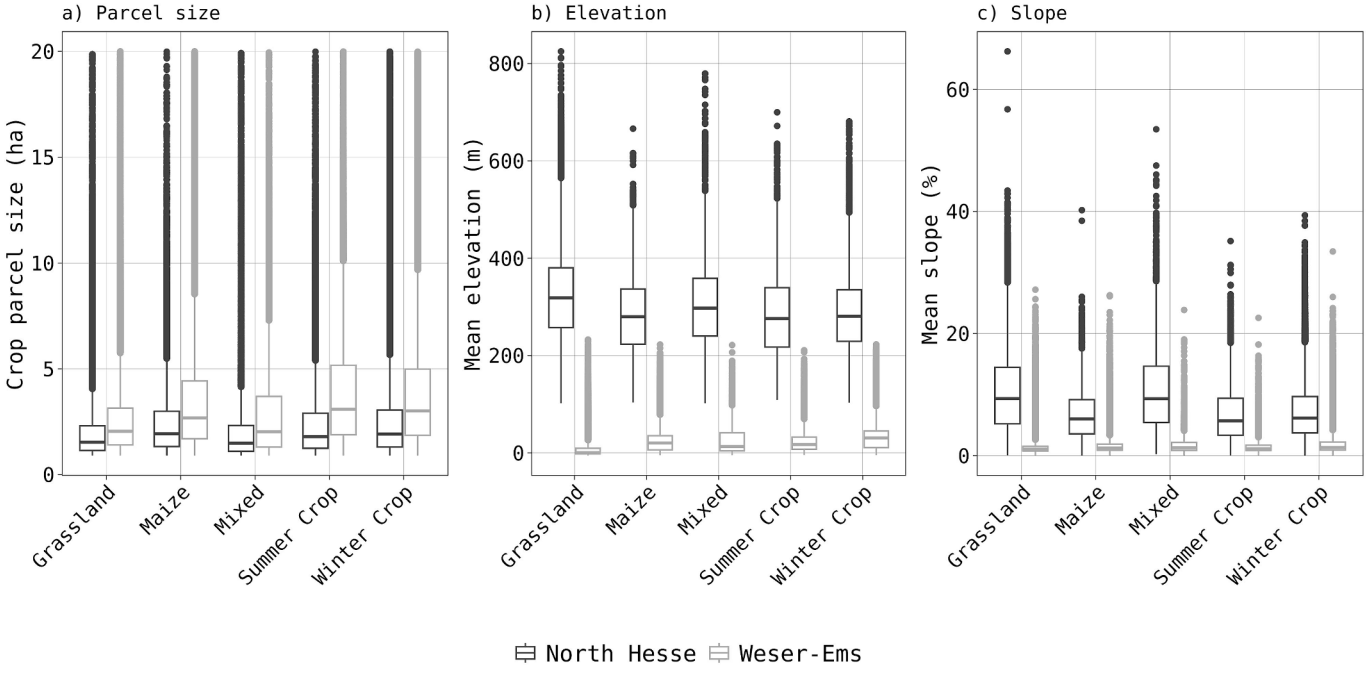

Fig. 11. Difference between a) Parcel size, b) mean elevation of the parcel, and c) mean slope of the parcel between the two regions and crop classes.

the newest developments, which are crucial for the full utilisation of transfer learning for multi-regional, past and present crop type mapping using RS time series.

5. Conclusions

This study is the first to quantify the performance loss of crop classification models trained on Landsat data when transferred to different spatial or temporal domains. It also investigates the potential causes of this performance loss. The study found that models trained on data from different spatial or temporal domains within the same country consistently underperform when transferred, regardless of the machine learning algorithm used. Therefore, further research employing multiple datasets that represent a multitude of spatial and temporal domains and applications of DA methods (shallow and deep) with novel DL models to evaluate the model transferability is recommended.

Based on the outcomes of this study following conclusions can be drawn:

- Crop-type mapping with sparse time series data from Landsat images is possible, and especially applicable for crop-type mapping tasks concerning the distant past.
- The model accuracies and macro F1 scores declined when the model was transferred to different spatial and/or temporal domains, and

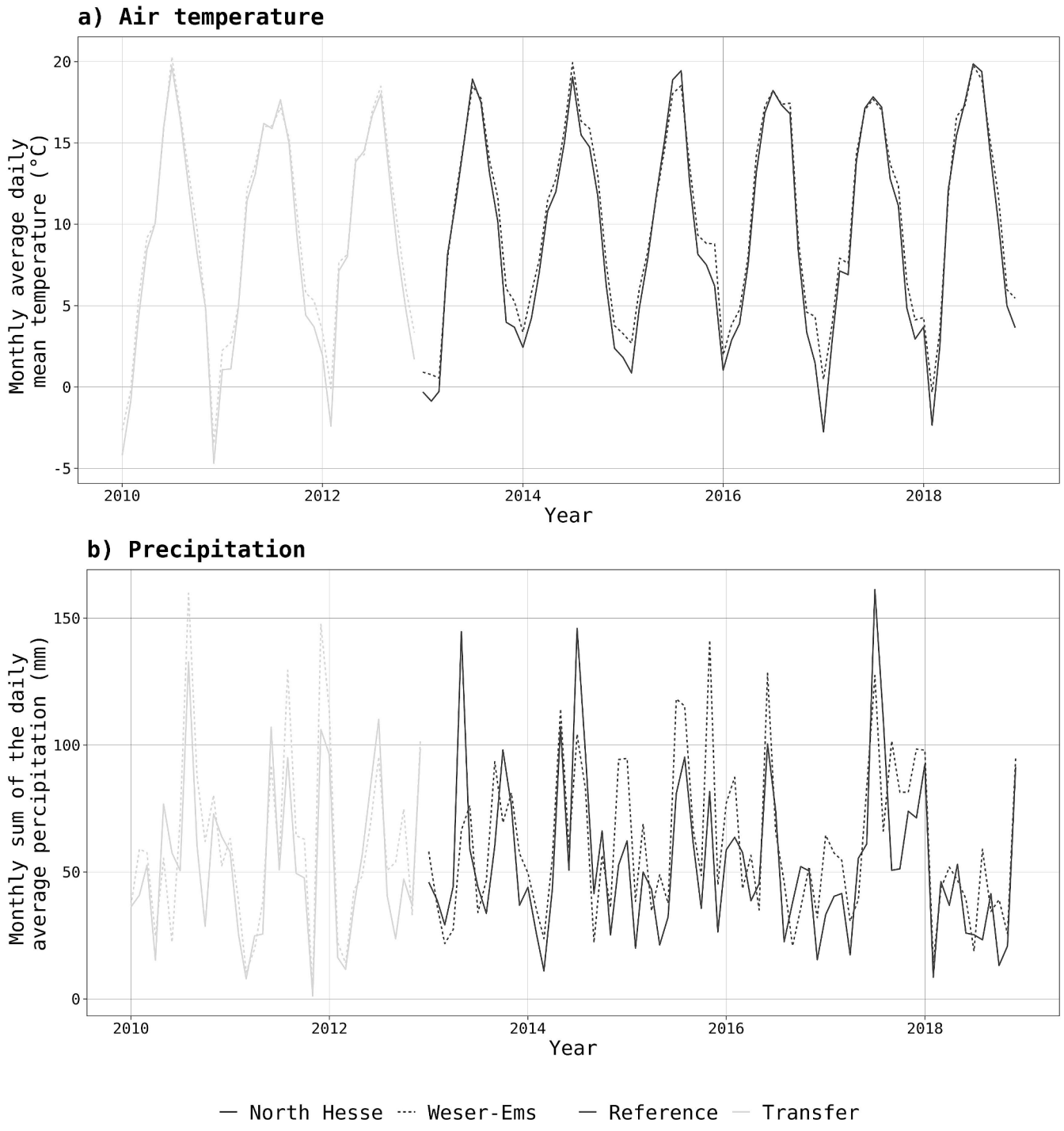

Fig. 12. Monthly a) temperature and b) precipitation in the two regions.

the order of model transferability loss was temporal (S1), spatial (S2) and spatial–temporal (S3).

- The two ML algorithms (RF and CNN) employed in this study showed no substantial differences in their performance.
- The observation percentage under the AOA was linked to the model accuracy, which could be a practical metric for understanding the new dataset's predictions using the trained model.
- The environmental, climatic and geographic data can be used to better elucidate the reasons for challenges in the model transferability scenarios.

6. Data and supplementary material

The employed Landsat data is freely available at respective databases (e.g., USGS earth explorer). The code is made available at https://github.com/jmatics/landsat-croptype-spatial–temporal-transferability. IACS data are not made available due to confidentiality reasons.

CRediT authorship contribution statement

Jayan Wijesingha: Conceptualisation, Data curation, Formal analysis, Methodology, Writing - original draft. **Ilze Dzene:** Project administration, Visualization, Writing – review & editing. **Michael**

Wachendorf: Funding acquisition, Resources, 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.

Acknowledgements

We would like to thank the Ministry of Agriculture of the Lower Saxony in Germany for providing historical agricultural land use data for the validation of the models.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF) under the project SYMOBIO 2.0 [grant number 031B1129A].

References

Ajadi, O.A., Barr, J., Liang, S.Z., Ferreira, R., Kumpatla, S.P., Patel, R., et al., 2021. Large-scale crop type and crop area mapping across Brazil using synthetic aperture radar and optical imagery. Int J Appl Earth Obs Geoinf 97, 102294. https://doi.org/10.1016/j.jag.2020.102294.

Asam, S., Gessner, U., González, R.A., Wenzl, M., Kriese, J., Kuenzer, C., 2022. Mapping Crop Types of Germany by Combining Temporal Statistical Metrics of Sentinel-1 and Sentinel-2 Time Series with LPIS Data. Remote Sens 14. https://doi.org/10.3390/rs14132981.

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., Hostert, P., 2022. Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. Remote Sens Environ 269. https://doi.org/10.1016/j.rse.2021.112831.

Braun, A.C., 2021. More accurate less meaningful? A critical physical geographer's reflection on interpreting remote sensing land-use analyses. Prog Phys Geogr 45, 706–735. https://doi.org/10.1177/0309133321991814.

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

Crawford, C.J., Roy, D.P., Arab, S., Barnes, C., Vermote, E., Hulley, G., et al., 2023. The 50-year Landsat collection 2 archive. Sci Remote Sens 8, 100103. https://doi.org/10.1016/j.srs.2023.100103.

Flood, N., 2014. Continuity of reflectance data between landsat-7 ETM+ and landsat-8 OLI, for both top-of-atmosphere and surface reflectance: A study in the australian landscape. Remote Sens 6, 7952–7970. https://doi.org/10.3390/rs6097952.

Frantz, D., Rufin, P., Janz, A., Ernst, S., Pflugmacher, D., Schug, F., et al., 2023. Understanding the robustness of spectral-temporal metrics across the global Landsat archive from 1984 to 2019 – a quantitative evaluation. Remote Sens Environ 298, 113823. https://doi.org/10.1016/j.rse.2023.113823.

Ghazaryan, G., Dubovyk, O., Löw, F., Lavreniuk, M., Kolotii, A., Schellberg, J., et al., 2018. A rule-based approach for crop identification using multi-temporal and multi-sensor phenological metrics. Eur J Remote Sens 51, 511–524. https://doi.org/10.1080/22797254.2018.1455540.

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.

Hooke, R.L.B., Martín-Duque, J.F., Pedraza, J., 2012. Land transformation by humans: A review. GSA Today 22, 4–10. https://doi.org/10.1130/GSAT151A.1.

Johnson, D.M., Mueller, R., 2021. Pre- and within-season crop type classification trained with archival land cover information. Remote Sens Environ 264, 112576. https://doi.org/10.1016/j.rse.2021.112576.

Kyere, I., 2020. Using optical satellite imagery to monitor and understand agricultural land-cover change. Universität Kassel. https://doi.org/10.17170/kobra-202006171350.

Kyere, I., Astor, T., Graß, R., Wachendorf, M., 2019. Multi-temporal agricultural land-cover mapping using single-year and multi-year models based on landsat imagery and IACS data. Agronomy 9, 13–15. https://doi.org/10.3390/agronomy9060309.

Kyere, I., Astor, T., Graß, R., Wachendorf, M., 2020. Agricultural crop discrimination in a heterogeneous low-mountain range region based on multi-temporal and multi-sensor satellite data. Comput Electron Agric 179. https://doi.org/10.1016/j.compag.2020.105864.

Levavasseur, F., Martin, L., Boros, L., Cadiou, J., Carozzi, M., Martin, P., et al., 2023. Land cover changes with the development of anaerobic digestion for biogas production in France. GCB Bioenergy 15, 630–641. https://doi.org/10.1111/gcbb.13042.

Ma, Y., Chen, S., Ermon, S., Lobell, D.B., 2024. Remote Sensing of Environment Transfer learning in environmental remote sensing. Remote Sens Environ 301, 113924. https://doi.org/10.1016/j.rse.2023.113924.

FAO, 2023. https://www.fao.org/faostat/en/#data (accessed January 18, 2023).

Meyer H, Milà C, Ludwig M, Linnenbrink J. CAST: "caret" Applications for Spatial-Temporal Models 2023.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol Evol 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Montero, D., 2021. eemont: A Python package that extends Google Earth Engine. J Open Source Softw 6, 3168. https://doi.org/10.21105/joss.03168.

Mueller, L., Schindler, U., Behrendt, A., Eulenstein, F., Dannowski, R., 2007. The Muencheberg Soil Quality Rating (SQR). Muencheberg.

Müller, H., Haberlandt, U., 2018. Temporal rainfall disaggregation using a multiplicative cascade model for spatial application in urban hydrology. J Hydrol 556, 847–864. https://doi.org/10.1016/j.jhydrol.2016.01.031.

National Climatic Data Center. Global Surface Summary of the Day - GSOD 2021. https://www7.ncdc.noaa.gov/CDO/cdosubqueryrouter.cmd (accessed March 17, 2021).

Niedersächsisches Ministerium für Ernährung Landwirtschaft und Verbraucherschutz. Die niedersächsische Landwirtschaft in Zahlen. 2022.

Nyborg, J., Pelletier, C., Assent, I., Nyborg, J., Pelletier, C., Assent, I., et al., 2022a. Generalized Classification of Satellite Image Time Series with Thermal Positional Encoding. CVPR, New Orleans, p. 12.

Nyborg, J., Pelletier, C., Lefèvre, S., Assent, I., 2022b. TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. ISPRS J Photogramm Remote Sens 188, 301–313. https://doi.org/10.1016/j.isprsjprs.2022.04.018.

Ofori-Ampofo, S., Pelletier, C., Lang, S., 2021. Crop type mapping from optical and radar time series using attention-based deep learning. Remote Sens 13, 1–17. https://doi.org/10.3390/rs13224668.

OpenTopography, 2013. 2023,. https://doi.org/10.5069/G9445JDF accessed May 15.

Orynbaikyzy A, Gessner U, Conrad C. Spatial Transferability of Random Forest Models for Crop Type Classification Using Sentinel-1 and Sentinel-2. Remote Sens 2022;14. DOI: 10.3390/rs14061493.

Pelletier, C., Webb, G.I., Petitjean, F., 2019. Temporal convolutional neural network for the classification of satellite image time series. Remote Sens 11, 1–25. https://doi.org/10.3390/rs11050523.

Peng, J., Huang, Y., Sun, W., Chen, N., Ning, Y., Du, Q., 2022. Domain Adaptation in Remote Sensing Image Classification: A Survey. IEEE J Sel Top Appl Earth Obs Remote Sens 15, 9842–9859. https://doi.org/10.1109/JSTARS.2022.3220875.

Pérez-Hoyos, A., Rembold, F., Kerdiles, H., Gallego, J., 2017. Comparison of global land cover datasets for cropland monitoring. Remote Sens 9. https://doi.org/10.3390/rs9111118.

Rusňák, T., Kasanický, T., Malík, P., Mojžiš, J., Zelenka, J., Sviček, M., et al., 2023. Crop Mapping without Labels: Investigating Temporal and Spatial Transferability of Crop Classification Models Using a 5-Year Sentinel-2 Series and Machine Learning. Remote Sens 15, 3414. https://doi.org/10.3390/rs15133414.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical Satellite Time Series Classification. ISPRS J Photogramm Remote Sens 169, 421–435. https://doi.org/10.1016/j.isprsjprs.2020.06.006.

See, L., Fritz, S., You, L., Ramankutty, N., Herrero, M., Justice, C., et al., 2015. Improved global cropland data as an essential ingredient for food security. Glob Food Sec 4, 37–45. https://doi.org/10.1016/j.gfs.2014.10.004.

Song, X.-P., 2023. The future of global land change monitoring. Int J Digit Earth 16, 2279–2300. https://doi.org/10.1080/17538947.2023.2224586.

Suck, R., Bushart, M., Hofmann, G., Schröder, L., 2014. Karte Der Potentiellen Natürlichen Vegetation, vol. Blatt 13. Bundesamt für Naturschutz, Bonn.

Vergara, F., Lakes, T., 2019. Maizification of the Landscape for. Biogas Production? 16, 1–27.

Wagner, B., 2011. Spatial analysis of loess and loess-like sediments in the Weser-Aller catchment (Lower Saxony and Northern Hesse, NW Germany). E G Quat Sci J 60, 27–46. https://doi.org/10.3285/eg.60.1.02.

Wetterdienst, D., 2018. Klimareport Niedersachsen. Deutscher Wetterdienst, Offenbach Am Main.

Wu, Q., 2020. geemap: A Python package for interactive mapping with Google Earth Engine. J Open Source Softw 5, 2305. https://doi.org/10.21105/joss.02305.

Wu, B., Zhang, M., Zeng, H., Tian, F., Potgieter, A.B., Qin, X., et al., 2023. Challenges and opportunities in remote sensing-based crop monitoring: a review. Natl Sci Rev 10, 17. https://doi.org/10.1093/nsr/nwac290.

Xu, Y., Ma, Y., Zhang, Z., 2024. Self-supervised pre-training for large-scale crop mapping using Sentinel-2 time series. ISPRS J Photogramm Remote Sens 207, 312–325. https://doi.org/10.1016/j.isprsjprs.2023.12.005.

Yuan, Y., Lin, L., 2021. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. IEEE J Sel Top Appl Earth Obs Remote Sens 14, 474–487. https://doi.org/10.1109/JSTARS.2020.3036602.