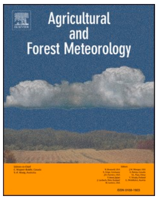
Agricultural and Forest Meteorology 341 (2023) 109652

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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



Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation

Rhorom Priyatikanto [a,b,*], Yang Lu [a], Jadu Dash [a], Justin Sheffield [a]

[a] *School of Geography and Environmental Science, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom*
[b] *Research Center for Space, National Research and Innovation Agency, Bandung 40173, Indonesia*

ARTICLE INFO

Keywords:
Yield prediction
Machine learning
Generalisability
Transferability
Domain adaptation

ABSTRACT

Modern problems in agricultural management require non-traditional solutions, one of which is by utilizing domain adaptive machine learning models for crop yield prediction which are able to perform reliably in different temporal or spatial domains. However, most studies have focused on the application of domain adaptation to classification tasks such as crop type identification, while the application to regression tasks such as crop yield prediction have been limited. In this study, we explore the generalisability and transferability of ordinary Deep Neural Network (DNN) and domain adaptive neural network models created using three domain adaptation algorithms, namely Discriminative Adversarial Neural Network (DANN), Kullback-Leibler Importance Estimation Procedure (KLIEP), and Regular Transfer Neural Network (RTNN). These three algorithms represent feature-based, instance-based, and parameter-based domain adaptations, respectively. Maize yield records, weather variables, and remotely sensed features from 11 states in the US corn belt acquired in 2006–2020 were compiled and segregated into classes according to temporal (year) and spatial characteristics (annual growing degree days [GDD], vapor pressure deficit [VPD], soil organic content [SOC], and green chlorophyll vegetation index/GCI). We found that models trained using datasets from temperate regions with medium-high GDD and moderate VPD perform well whereas SOC does not significantly affect the generalisability. It is not advisable to train models with datasets constrained by GCI as this feature correlates significantly with the maize yield, and adaptation between two domains that rarely intercept will not work well. We also demonstrate that Kullback-Leibler divergence computed using features from source and target domains can be used to justify the feasibility of domain adaptation. Based on the divergence, a model trained in the US (or another region with sufficient data) is expected to work reliably in other regions through domain adaptation, especially feature-based adaptation.

1. Introduction

The fast-growing human population, which is expected to reach 9.7 billion in 2050 (UN, 2019), requires a significant increase in global food production. For instance, to meet the demand for cereals in 2050, the global production must increase by 25%−70% compared to the actual production in 2014 (Hunter et al., 2017). This figure is in line with the expected rise of the total global food demand from van Dijk et al. (2021). By considering plausible socioeconomic pathways, van Dijk et al. (2021) estimated an increase of 35% to 56% in food demand, from 2010 to 2050. This poses a grand challenge in maintaining food security under a changing climate (Kang et al., 2009; Kumar, 2016). Intensification and expansion of agricultural activities are possible solutions to increase food production. Among others, timely crop monitoring and accurate estimation and prediction of crop yield plays an important part in developing policy to maintain food security and close the gap between attained and potential yield (Mueller et al., 2012). Nowadays, these approaches include the utilization of multispectral or even hyperspectral remote sensing data in tandem with relevant physical parameters and in-situ measurements to develop data-driven models of crop yield (Azzari et al., 2017; Yang et al., 2021; Yoosefzadeh-Najafabadi et al., 2021; Vergopolan et al., 2021). Simple linear regression (Becker-Reshef et al., 2010; Qader et al., 2018), multiple linear regression (Gonzalez-Sanchez et al., 2014), and machine learning (Johnson, 2016; You et al., 2017; Kang et al., 2020) have been used to estimate and predict yields of different crop types using remotely sensed and other data, each

* Corresponding author.
E-mail address: rhorom.priyatikanto@brin.go.id (R. Priyatikanto).

https://doi.org/10.1016/j.agrformet.2023.109652
Received 2 September 2022; Received in revised form 30 July 2023; Accepted 4 August 2023
Available online 9 August 2023

with its own advantages and limitations (Duncan et al., 2015). At one end of the list of approaches, linear regression is commonly used as the basis of yield prediction due to its simplicity and frugality in terms of multidimensional data requirements. At the other end, various machine learning techniques alternatively generate models by accounting for non-linear relationships between multiple variables with crop yield as the dependent variable. Supported by increased computing resources, machine learning techniques have become an increasingly popular approach (Chlingaryan et al., 2018; van Klompenburg et al., 2020).

There are a number of recent studies that use remote sensing data and machine learning techniques for predicting yields (Park et al., 2018; Zhao et al., 2020; Vergopolan et al., 2021) with an aim to derive accurate yield estimates at county/district level. These studies are typically performed by using remotely sensed variables in combination with weather data with a typical resolution of ~1 km. For example, Kang et al. (2020) built machine learning models for predicting maize yield in the US corn belt and achieved well-performing models with 9% mean absolute percentage error. Other studies have focused on estimating crop yield at higher spatial resolution, e.g. by incorporating 10-m resolution data from Sentinel 2 and field level crop yield data (e.g., Jin et al., 2019). So far, the performance of published models are mostly represented by coefficients of determination (R2) that range from 0.5 to 0.9 with the note that the values depend on the training and validation data utilized. In general, accommodating more input data generally increases the performance of the model (Vergopolan et al., 2021) while limiting the dimension of input data may yield underperforming models, even compared to simple linear regression as the benchmark (Meroni et al., 2021).

The performance of machine learning models is highly reliant on the quality and volume of input data, hence previous studies have been predominantly performed in data-rich regions as the source domain. The transferability of established models to a different target region is subject to influence from the divergence between source and target domains. Some studies have mentioned this matter as the limitation of machine learning models. Loss of accuracy may occur when the models encounter conditions at extreme ends of the training data or beyond, such as different climatic conditions in the future that severely affecting the crop yield (Jeong et al., 2016). Traditional machine learning models trained using a specific dataset may require re-training to gain acceptable performance when dealing with another dataset (Pan and Yang, 2009; Xu et al., 2021). In reality, the re-training process cannot be performed since the target domain may be lacking labelled data or the re-training may be costly. This hinders the application of well-trained machine learning models to often food-insecure regions such as Africa and South Asia where reliable data is not always available. Transfer learning through domain adaption can potentially improve machine learning performance in this context, though the application to crop yield prediction is in its infancy.

Domain adaptation is a subset of transfer learning methods which aim to transfer knowledge established during the training process in one source domain to improve the predictive capability in a different target domain. Domain adaptation is regarded as transductive transfer learning where the model deals with data from domains with slightly different distributions, but performs the same tasks like classification or regression in another domain (Pan and Yang, 2009). There have been many applications of domain adaptation for classification tasks based on remotely sensed data (Tuia et al., 2016; Teng et al., 2020). Through domain adaptation, the impact of domain shifts caused by the utilization of nonidentical sensors, data acquisition in different conditions, or the spatial variations, is minimized to improve prediction (Pacifici et al., 2014;Walker et al., 2012; Li et al., 2021). In the field of remote sensing, domain adaptation can be performed through: (i) selection of invariant features, (ii) adaptation of the data distribution, and (iii) adaptation of the classifier/regressor (Tuia et al., 2016). In the first approach, the data from source and target domains are compared until features that are less affected by domain shift can be identified (Izquierdo-Verdiguier et al., 2013). In the second approach, instance-based adaptation can be used to re-weight the instance in both source and target data such that the divergence between them is suppressed (Pan and Yang, 2009). In the third approach, the parameters of the model(s) trained using the source domain are tweaked according to the characteristics of the target domain. For this purpose, a subset of labelled data from the target domain is required.

Apart from application to classification tasks, applications of domain adaptation to regression tasks that invoke remote sensing data is limited. For example, Wang et al. (2018) presented the application of deep transfer learning for crop yield prediction with the transfer from Argentinean crops as the source domain and Brazilian crops as the target. More recently, Ma et al. (2021) reported the implementation of an adaptive domain adversarial neural network (ADANN) for crop yield prediction in the US corn belt where the study area was divided into two major ecosystem regions, namely eastern temperate forests and great plains. In that study, the domain adaptation scheme projected the features from both source/training and target/test domains into the same subspace through adversarial learning (Ganin et al., 2016). The transferability experiments were performed by training the model using projected data from one region and testing using data from another domain. They demonstrated that models with transfer learning through ADANN outperform Random Forest and Deep Neural Network baseline models. Moreover, spatial variation of the prediction error is obvious when direct transfer is applied while transfer learning reduces this kind of variation.

The error generated during model testing using the target domain depends on the empirical source error and the divergence between the source and target domains (Ben-David et al., 2010). The divergence itself can be estimated using statistical measures like Kullback-Leibler divergence (Kullback and Leibler, 1951; Perez-Cruz, 2008). Based on this theorem, we can justify the conditions (in terms of domain divergence) that enable domain adaptation to perform well. In domain adaptation, it is essential to know when, what, and how to transfer knowledge (Tuia et al., 2016). The first aspect is related to the conditions on which the domain adaptation perform sufficiently well in improving the accuracy of machine learning models. The latter two aspects are related to the methods or approaches to take while performing adaptation. Among the three families of domain adaptation methods mentioned before, one may be good for a specific case while others may work better in other situations. The preference on what and how to adapt depends on the specific problem considered (Tuia et al., 2016).

In this study, we explore the generalisability and transferability of machine learning models and also the capability of domain adaptation to alleviate the problem of deteriorating performance when machine learning is implemented on different domains, specifically focusing on crop yield prediction. By using data from the US corn belt which is segregated temporally and spatially, we address issues related to what, how, and when to perform domain adaptation for crop yield prediction. Three domain adaptation algorithms representing feature-based, instance-based, and parameter-based approaches are evaluated, with an ordinary deep neural network without transfer learning used as the baseline model. The anticipated results are important because state-of-the-art machine learning models need to tackle real data with various characteristics. Through domain adaptation (or other kinds of transfer learning techniques), models trained in a data rich country or region are expected to be applicable to the other countries or regions with limited training data. This would improve crop yield prediction in many data-poor regions, which often have the most need for research into improving crop production and food security.

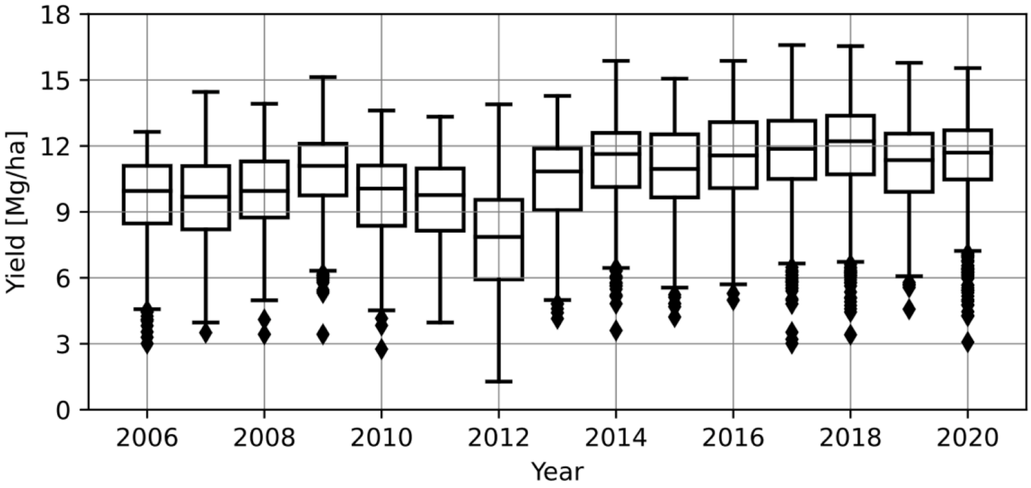


Fig. 1. The boxplot of the district-level maize yields from the study area as a function of year. An increasing trend of yields can be seen beside the drop in 2012.

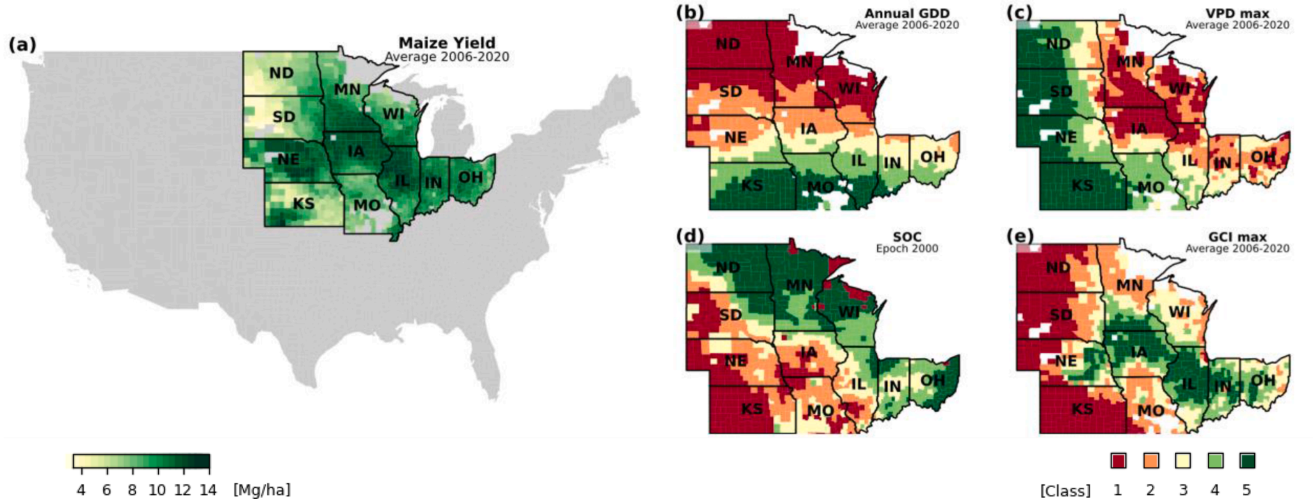


Fig. 2. Area of study includes 11 states in the US corn belt (a). Spatially, this area can be categorised into classes according to the annual Growing Degree Days (b), maximum Vapor Pressure Deficit in August (c), Soil Organic Content (d), and the maximum Green Chlorophyll Index (e). The categorisations shown in the maps are based on the average value in 2006–2020 while the actual categorisations on yearly data may vary slightly.

2. Data and methods

2.1. Area of study

The United States is the top maize-producing country in the world with a total production of 383 million tonnes in 2021 and record-breaking productivity of 11.9 tonnes per hectare (USDA, 2022). The US corn belt is an ideal area for experimentation using machine learning because of wealth of data including a reliable crop mask for most of the corn production regions since 2006 and abundant remote sensing and climate data. We acquired crop yield records from eleven states (Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin) from the National Agricultural Statistics Service of the United State Department of Agriculture (USDA/NASS), which has provided tabulated county-level maize yield since the 1980s. The Cropland Data Layers product (Boryan et al., 2011) was used to indicate maize crop activity in a certain year. We used crop yield records from 2006 to 2020 to build the training and testing datasets for our experiments (see Fig. 1). The complete dataset contains 10, 546 crop yield records (all years combined) from 845 counties.

Following the work of Ma et al. (2021), the area of study was segregated into five different classes according to a defined categoriser where generalisability and transferability experiments can be performed using data from those classes. Categorisers used are year, annual Growing Degree Days (GDD), Vapor Pressure Deficit (VPD), Soil Organic Content (SOC), and the maximum Green Chlorophyll Index (GCI). These features were selected as categorisers because they show a gradual pattern across the area of study (see Fig. 2). To be noted that the year was used as categoriser but it was not used as feature in the modeling. Year can be regarded as the standard categoriser where the machine learning models are usually tested using data from different years (Ju et al., 2021). GDD correlates with the geographical latitude while the VPD map shows a gradual change from east to west. SOC almost resembles the ecological system defined by the US Environmental Protection Agency (EPA, 2001) while the GCI represents the effectiveness of agricultural activities in the region. Further descriptions of these categorisers are provided in Table 1.

2.2. Remote sensing and weather data

Based on the review of van Klompenburg et al. (2020), there are seven groups of features that are commonly used in the machine learning models for crop yield prediction, namely soil information (type, moisture, content, etc.), solar radiation information (incoming

Table 1
Description of features used as the predictors.

Feature	Formula	Remarks
Green Chlorophyll Index (GCI)	$\frac{B_2}{B_4} - 1$	MODIS Nadir Bidirectional Reflectance Distribution Function Adjusted Reflectance (MCD43A4.006) Resolution: 500 m, daily Source: 10.5067/ MODIS/MCD43A4.006 B1: red (620–670) B2: NIR (841–876 nm) B3: blue (459–479 nm) B4: green (545–565 nm) B6: SWIR (1628–1652 nm)
Enhanced Vegetation Index (EVI)	$\frac{2.5(B_2 - B_1)}{(B_2 + 6B_1 - 7.5B_3 + 1)}$	
Normalized Difference Water Index (NDWI)	$\frac{B_2 - B_6}{B_2 + B_6}$	
Fraction of Photosynthetically Active Radiation (FAPAR)		MODIS Leaf Area Index product (MOD15A2H). Resolution: 500 m, 8 days Source: 10.5067/ MODIS/MOD15A2H.006
Daytime Land Surface Temperature (LST)		MODIS Terra Land Surface Temperature and Emissivity (MOD11A1.006) Resolution: 1000 m, daily Source: 10.5067/ MODIS/MOD11A1.006
Average Temperature (Tmean)		Parameter-elevation Relationships on Independent Slopes Model (PRISM, Daly et al., al.,2008, 2015) Resolution: 4000 m, daily
Growing Degree Date (GDD)	$\sum \min(\max(Tmean, 10), 30) - 10$	
Vapor Pressure Deficit (VPD)		
Soil Organic Content (SOC)		SoilGrids (Hengl et al., 2017) Resolution: 250 m, single epoch: 2000
Elevation		SRTM digital elevation model Resolution: 90 m, single epoch: 2000 Source: https://srtm.csi.cgiar.org

shortwave radiation), weather (humidity, precipitation, temperature etc.), nutrients (nitrogen, magnesium, etc.), crop information (type, density, etc.), crop growth parameters (vegetation indices, canopy cover, etc.), and field management practices. Not all the features have a significant effect on crop yield prediction. Some features have higher correlation with annual crop yield while others may only have an indirect association to the crop yield. Some features are time-invariant while others are dynamically changing over time. According to the analyses by Johnson et al. (2016) and Kang et al. (2020), dynamic features such as vegetation indices and land surface temperature obtained in the mid-season correlate well with annual crop yield. Consequently, these features can be good predictors for within-season crop yield prediction. Different regions may exhibit different phenology of corn, but for most of the US corn belt, the season peaks in August (Johnson, 2014; Kang et al., 2020). In this study, monthly aggregate values of the dynamic variables acquired in August are selected as the predictive features. In contrast to some studies that utilised tens of features derived from time-series data with a wider range of time, we demonstrate that the use of peak season data is enough to produce a good crop yield prediction.

Based on the Pearson's correlation coefficient with the crop yield and the feature importance metric computed during the training of machine learning models, among twenty features (see Table A1 in the appendix) we selected the best ten features to be included in the experiments. Variables with high correlations to the crop yield and high importance scores (mean decrease of accuracy) were selected. We also performed variable reduction by dropping similar variables with high intercorrelation. For instance, Normalised Difference Vegetation Index (NDVI) well correlates with EVI while its importance is less than EVI. We ended up with GCI, Enhanced Vegetation Index (EVI), Fraction of Photosynthetically Active Radiation (FAPAR), Normalized Difference Water Index (NDWI), Daytime Land Surface Temperature (LST), mean temperature (Tmean), GDD, VPD, SOC, and elevation. Vegetation indices and LST were derived from remote sensing data acquired using the Moderate Resolution Imaging Spectroradiometer (MODIS). The annual GDD was calculated from the daily average temperature from the Parameter-elevation Relationships on Independent Slopes Model (PRISM, Daly et al., 2008, 2015). PRISM is a product of Climatologically-Aided Interpolation with the input of weather parameters from ground stations and 4-km resolution outputs. The maximum daily VPD is also provided in the PRISM. Static information regarding soils, especially SOC was obtained from SoilGrids (Hengl et al., 2017) which contains global predictions of several soil properties derived from soil profiles from around the world and some remote sensing data with epoch 2000. SoilGrids provides soil properties in seven standard depths (0 to 200 cm), but we only use the surface properties (0-cm depth). Lastly, the elevation data was extracted from the Shuttle Radar Topography Mission (SRTM) digital elevation dataset. Descriptions of these features including the sources and resolutions are provided in Table 1. As an additional information, the full list of features including the ones not selected is provided in Table A1.

As a matter of outlook (section 3.4), we also acquired features from other corn producing states in the US and some selected regions outside the US. For the regions outside the US where the PRISM dataset is not available, we used TerraClimate (Abatzoglou, 2013) as the source of weather parameters. Additionally, the crop mask is provided by the Global Food-Support Analysis Data (Teluguntla et al., 2015). The dynamic features (vegetation indices and weather parameters) are associated with the mid-seasons which are different across countries.

All remote sensing data and weather models listed in Table 1 were accessed and pre-processed in Google Earth Engine (Gorelick et al., 2017). The pre-processing stage includes quality assessment for daily remote sensing data, temporal aggregation, cropland masking, and spatial averaging over the counties. As mentioned before, we only selected the maximum values of the dynamic features acquired in August each year. We prefer the maximum over the average because the maximum values correlate more to the annual crop yield (Bolton and Friedl, 2013). Next, we employed the Cropland Data Layer (CDL, Boryan et al., 2011) to mask the maize-planting area. Due to crop rotation practices, it is necessary to use the year-specific CDL. Finally, both static and dynamic features were averaged over the county areas and extracted into tabulated values for experiments using machine learning.

2.3. Machine learning models

We used Deep Neural Network (DNN) as the baseline model where the dataset from the source domain ($X_S \in X$) is passed through a series of densely-connected layers with a hundred neurons at every layer. At the end of the network, there is an output layer which is related to the crop yield as the model output ($X_S \in Y$). In this context, X is the input feature space while Y is the crop yield. The model learns the distribution of D (X_S, Y_S) and predicts the relevant distribution for the target domain D (X_T, Y_T) which is not always similar to the distribution in the source domain. Inside the DNN, a non-linear activation function is used to select or to weight neurons during the transition between layers such that a certain neuron may be activated or deactivated in the network. Rectified linear unit (ReLU) is a popular choice of activation function considering its simplicity that enables faster learning processes (Lecun

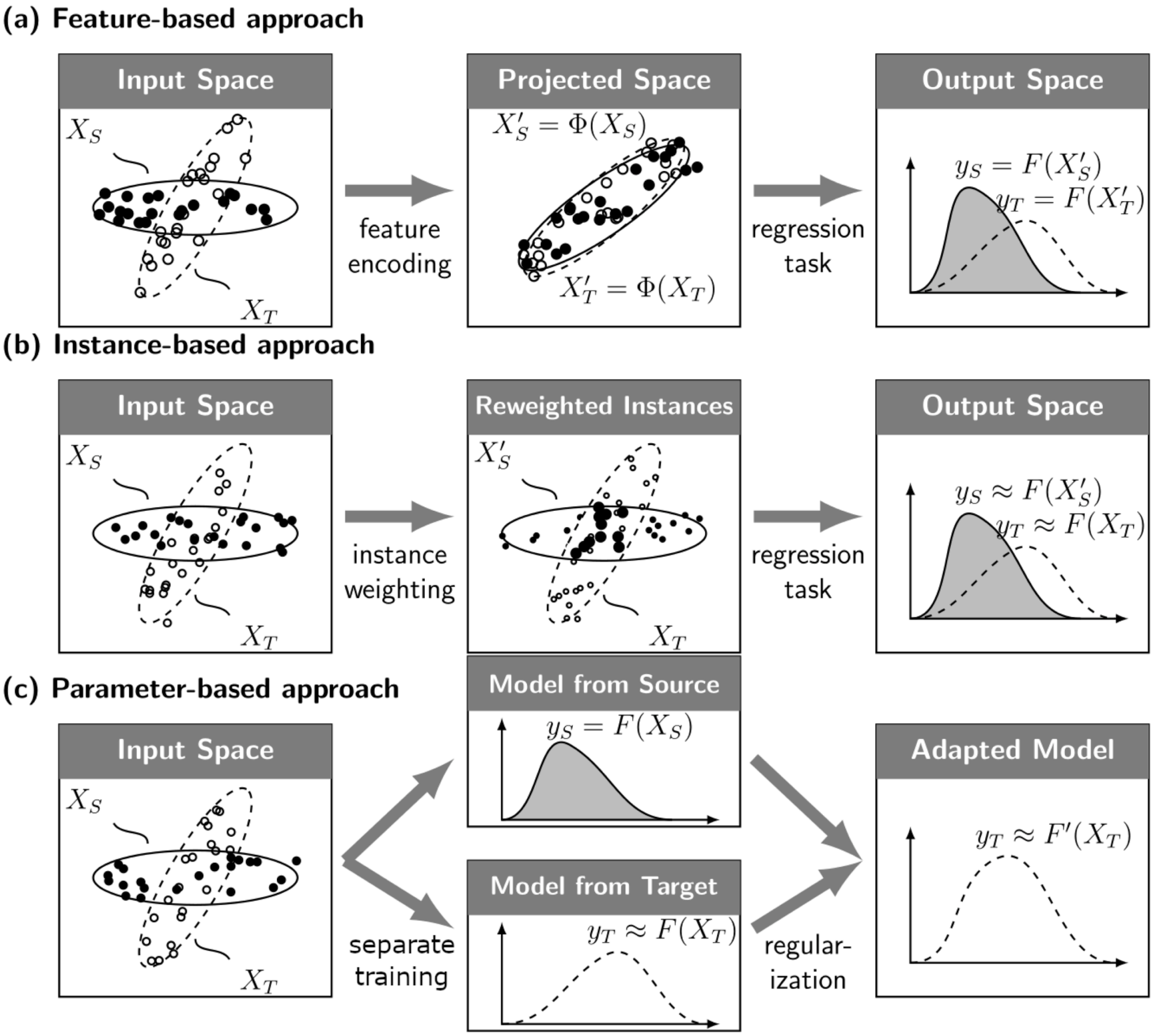


Fig. 3. Schematic illustrations of (a) feature-based, (b) instance-based, and (c) parameter-based domain adaptations.

et al., 2015; Glorot et al., 2011). The parameter or weight of every neuron at every layer is adjusted during the training processes such that the loss function is minimized.

In practice, we employed DNN with an input layer, six hidden layers, and one output layer which was created using the Tensorflow version 2.7 package in Python (Abadi et al., 2015). Each hidden layer consists of 100 neurons activated using ReLU as the activation function. The model was trained in 50 epochs with the Adam algorithm (Kingma and Ba, 2014) as the stochastic gradient descent method to optimize the model. A constant learning rate of 0.001 was set during the optimization. This configuration was selected among different hyperparameter sets evaluated through systematic training and testing with a small dataset. In general, deeper neural networks do not always provide better results while overtraining in many epochs without proper batch normalization may yield overfitting.

For the domain adaptation, we used ADAPT (Awesome Domain Adaptation Python Toolbox) package developed by de Mathelin et al. (2021). This toolbox contains three major classes of domain adaptation which are feature-based, instance-based, and parameter-based adaptations (see Fig. 3). Feature-based domain adaptation can be regarded as representation learning where the features from both source and target domains are projected to a common space such that the source and target become indistinguishable to a certain degree. Among several alternative algorithms, we selected Discriminative Adversarial Neural Network (DANN, Ganin et al., 2016) to represent feature-based approaches. Initially, DANN was developed and tested for classification tasks, but it has been used in some studies related to regression tasks, including maize yield prediction by Ma et al. (2021). The architecture of DANN consists of three major parts which are the deep feature extractor or encoder, the deep label predictor or task layers, and the domain classifier. In the common machine learning algorithm, an encoder transforms categorical data into numerical form while in our DANN, the encoder encapsulates an encoder function ($\Phi(X)$) that extracts and projects the numerical features into a certain space such that source and target domains cannot be distinguished by the domain classifier ($D(X)$).

Parallel with that, projected features from the source domain are passed through the task layers ($F(X)$) up to the output layer such that the loss function can be evaluated. The gradient reversal layer, which connects the domain classifier and encoder, establishes the adversarial learning part of the DANN (see Fig. 4). The overall learning process can be performed using standard back propagation and stochastic gradient descent methods such that the DANN scheme can be implemented with common deep learning packages (Ganin et al., 2016). In the case of domain adaptation with unlabelled data from the target domain, the loss incurred in the prediction task for the source needs to be minimized together with the divergence between features from the source and

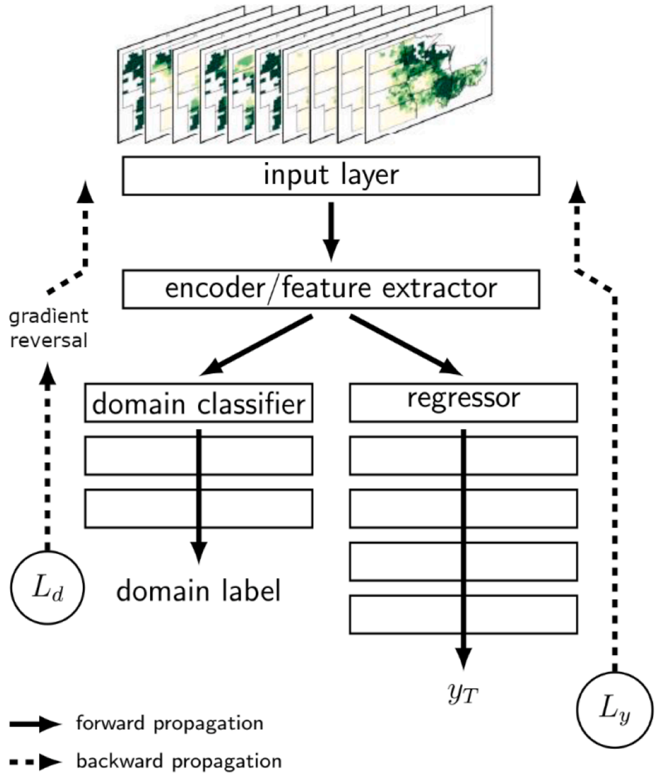


Fig. 4. Illustration of the DANN that consists of single layer encoder, three layers domain classifier, and five layers of regressor. Comparable DNN only contains input and regressor layers.

target domains. Following this idea, the optimization problem in DANN is

$$\underset{F,\Phi}{\text{argmin}}\,(\mathrm{L}(\mathrm{F},\Phi,\mathrm{X_S})-\lambda\,\mathrm{R}(\mathrm{D},\Phi,\mathrm{X_S},\mathrm{X_T}))$$

where Φ, *F*, and *D* are encoder, task, and discriminator functions, *L* is the task loss function, while *R* is regularization function.

$$R=\log(1-D(\Phi,X_S))+\log(D(\phi,X_T))$$

In DANN, λ tunes the trade-off and convergence between domains and the prediction loss. When the features from the source and target domains converge or are indistinguishable in the new feature space, we can expect a model with better performance when implemented in the target domain. We experimented with DANN consisting of an input layer, single layer as encoder, five hidden layers for prediction task, three hidden layers for discriminating domains and one output layer for both task and discriminator. Every hidden layer contains 100 neurons with ReLU as the activation function such that the DANN basically has a similar configuration compared to the baseline DNN defined before. This relatively shallow neural network is in contrast with the one defined in Ma et al. (2021), where the encoder has more layers than the remaining, though the depth is similar. DANN with less encoder layers can be perceived as domain adaptation with simpler non-linear feature projection. Instead of using variable or adaptive values of λ as in Ma et al. (2021), we used a fixed $\lambda = 1.0$ in our experiments.

In the next approach, instance-based adaptation was performed by recalculating the weight for every instance or data point considering the difference between source and target domains. More weights are assigned to instances from the source domain that are located in the feature space intercepting with the target domain such that the model can perform better in the target domain. Normally, the weight is defined as the ratio between densities estimated from source and target domains, but the empirical density estimation in multidimensional space becomes a cumbersome task. Sugiyama et al. (2007) proposed an alternative way of instance weighting through Kullback-Leibler Importance Estimation Procedure (KLIEP) where the weights are directly estimated based on the instances from source and target domains while the convergence between the two domains after re-weighting is measured using Kullback-Leibler divergence (Kullback and Leibler, 1951). In KLIEP, the main optimization problem is

$$\underset{\alpha_i}{\text{argmax}}\sum_{X_T}\log\left(\sum_{X_T}\alpha_i K_\sigma\right)$$

subject to

$$\sum_{X_S}\left(\sum_{X_T}\alpha_i K_\sigma\right)=n_S$$

such that the new weight becomes

$$w(X_S)=\sum_{X_T}\alpha_i K_\sigma$$

Here, x are the instances either from source (X_S) or target (X_T) domain, α are the basis functions coefficients, while K_σ are kernel functions with bandwidth of σ. The Gaussian kernel with $\sigma \in [0.001, 0.01]$ was used in our experiments. After the weights are estimated, the DNN model defined before can be implemented on the target dataset.

Lastly, we used Regular Transfer with Neural Network (RTNN, Chelba and Acero, 2006) as the representation of parameter-based domain adaptations. Different from previous approaches, RTNN is semi-unsupervised domain adaptation since it requires some labelled data from the target domain, e.g. crop yield records from the target dataset. In this approach, the models trained using source and target datasets are assumed to share parameters or a prior distribution of hyperparameters such that the domain adaptation becomes a regularization problem (Pan and Yang, 2009). RTNN works by fitting the neural network to the source dataset to optimize the following equation:

$$\underset{\beta}{\text{argmin}}\,||F(X_S,\beta)-Y_S||^2$$

where β represents possible parameters for the neural network F consisting of d layers. Then, the parameters for the target domain are obtained by solving the following

$$\underset{\beta=(\beta_1,\ldots,\beta_d)}{\text{argmin}}\,|F(X_T,\beta)-Y_T|^2+\sum_{i=1}^{d}\lambda_i\left|\beta_i-\beta_{s_i}\right|^2$$

The λ are the trade-off parameters that determine the generalisability and transferability of the domain adapted model. Intuitively, we expect that the performance of domain adaptation depends on the size and quality of the labelled target dataset supplied to the algorithm. If we have more labelled data that represents the general distribution of the target domain, we expect a better chance of adaptation. In contrast, the adaptation will not improve the model if the supplied target data is far from the ridgeline. In the experiments, we randomly selected 40 labelled target datasets (approximately 2% of the whole sample) and used fixed $\lambda = 1.0$ for all layers.

2.4. Experiment design

Previously, we mentioned that the complete dataset containing ~10,000 rows of data can be categorised into equal-size classes according to five categorisers (year, GDD, VPD, SOC, and GCI). Based on the year, we divided the data into eight classes each containing ~1400 data from 2 years of record. For the remaining categorisers, we binned the data into five equal-size classes (~2400 data values in each class). The generalisability experiments were performed by training the models (DNN, DANN, KLIEP, and RTNN) using data from a certain class (e.g.,

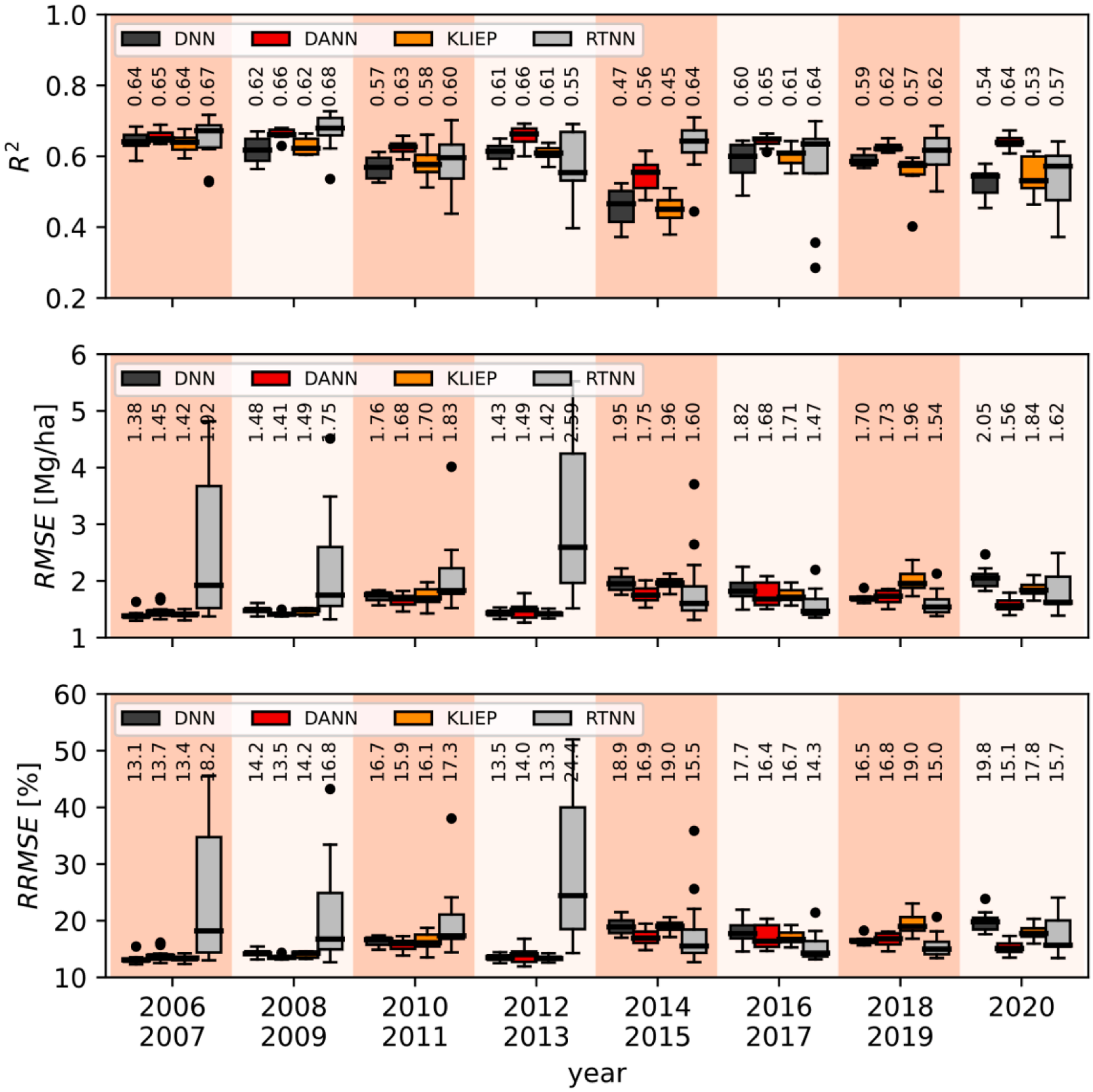


Fig. 5. Boxplot of coefficient of determinations (R^2, top), root mean square errors (RMSE, middle), and relative root mean square errors (RRMSE, bottom) of the models trained using data from selected years and tested on the remaining years. The median values are marked with thick horizontal lines in the boxplot and also printed on the panels.

years 2006–2007) as source domain and testing the models using data from the remaining classes (e.g., years 2009–2020). Meanwhile, transferability evaluations were performed on the models using data from a specific class (e.g., years 2008–2009) as the target. For each source and target pair, Kullback-Leibler divergence (D_{KL}) between features from those domains was calculated using the method of Perez-Cruz (2008). Basically, D_{KL} measures the distance between density $P(X_{\mathrm{S}})$ and $P(X_{\mathrm{T}})$ by evaluating the following equation

$$D_{KL} = \int p(x_S) \log \frac{p(x_S)}{p(x_T)} dx$$

The value of D_{KL} is nearly zero when both densities are similar ($P(X_{\mathrm{S}}) \approx P(X_{\mathrm{T}})$) and it grows as the densities diverge. In the method proposed by Perez-Cruz (2008), D_{KL} is estimated according to the cumulative density function or k-nearest-neighbours density estimation. This method is applicable for multidimensional data.

In our experiments, the divergences between source and target features categorised by years have typical $D_{\mathrm{KL}} \approx 6$, while other categorisations can produce D_{KL} from 3 to 22. In more detail, the divergence between the northernmost and the southernmost regions (categorised by the annual GDD) is about 19. The same extreme divergence is associated with the pair of driest and the most humid regions (according to the maximum VPD). The divergences between classes categorised by the SOC range from 3 to 17 while the divergences between GCI classes are slightly higher with the range of $4 - 22$.

For evaluating the model performance, we computed the coefficient of determination (R^2) and root mean square error (*RMSE*). We also normalised the RMSE by the average crop yield to obtain the relative root mean square error (RRMSE) in percentage. As a rule of thumb, an excellent model has a typical RRMSE of <10%, a good model has 10–20% RRMSE, a fair model has 20–30% RRMSE, while a model with more than 30% RRMSE is considerably poor. Considering the fact that the outputs of deep learning are subject to the stochastic variations in terms of performance scores, we repeated the training and testing procedures 16 times for each source and target domains pair. The statistical properties, especially the median score can be calculated using the resulted scores.

3. Results

3.1. *Inter-annual generalisability*

The first result from the generalisability experiments is associated with the models trained using data from certain years and tested on the data from the remaining years. Because the divergence between source and target domains categorised by year is relatively low, this result can be compared with the results reported in the literature, such as the works of Kang et al. (2020) and Ma et al. (2021). A more homogeneous distribution of features and yields across years is expected, unless a weather anomaly occurred in a certain year or there is an unaccounted factor that affected the crop yield.

Fig. 5 shows the performance scores achieved by the four models as a function of training years. In general, the median R^2 is around 0.60 with typical interquartile range (IQR) less than 0.05. A small declining trend in R^2 is observable especially for DNN and KLIEP, while a significant drop is obvious for models trained using data from 2014 to 2015. Anomalous condition caused by drought in 2012 (Lobell et al., 2014) does not produce significant decrease of generalisability of the models trained using data from that year. A plausible explanation for this result is that the drought affects the values of remotely sensed features in the same way as it affects the annual crop yield. Consequently, the anomaly does not affect the generalisability. Conversely, there may be other factors affecting the crop yield in 2014–2015 and the following years that are not accounted for in the models. Models trained using data before 2014 have a lower RMSE compared to the ones that were trained using data in 2014 and after (with 95% confidence interval), except for RTNN that shows the opposite pattern. Relatively low R^2 achieved by models trained using 2020 data are likely to be associated with the size of the dataset (approximately half of the other training datasets). From the probabilistic point-of-view, a smaller dataset has a lower capability to capture the characteristics of the general population. The behaviors and performances of data-driven models are dependent on the characteristics of the data utilized in training and testing processes (Reichstein et al., 2019).

Similar patterns can be observed when we evaluate the RMSE and RRMSE as the measure of generalisability (Fig. 5 middle and bottom). The median RMSE are more than 1.4 Mg/ha with typical IQR of less than 0.5 Mg/ha. An increasing trend of RMSE is clear for the DNN without domain adaptation. The IQRs tend to be wider since 2014, except for the RTNN where the performance scores are governed by randomly selected labelled data from the target domain. The same patterns are observed in the RRMSE plot as the variation of the average yields in the defined target domains is negligible. For most of the cases, the median RRMSEs stay under 20% indicating good generalisability.

Domain adaptive models do not always outperform the baseline DNN model. In terms of R^2, models with KLIEP are comparable to the DNN while KLIEP tends to produce higher errors. If we assume that the categorization by year does not produce highly divergent source and target domains, then the source domain in the generalisability experiment can be regarded as a subset of the target domain. Consequently, recalculating the weights of the instances as in KLIEP does not improve the generalisabililty score significantly. Next, RTNN models which were supplied with some labelled data from target domains tend to perform better than the DNN, though the performance scores are strongly affected by the selection of the labelled data for regularization. This is indicated by the wide IQR plotted in Fig. 5. Lastly, DANN models have a higher generalisability score compared to the DNN, both in R^2 and RMSE.

In addition to the above analyses, we can also compare the scores achieved by the models in this study with the same measures in the literature. From the local experiments presented in Ma et al. (2021), the R^2 for DNN are within $0.55-0.81$ while the RMSE range from 0.95 to 1.26. For the DANN, the scores are $R^2 = [0.47, 0.81]$ and RMSE $= [1.00, 1.65]$. Lastly, ADANN performs best with $R^2 = [0.62, 0.85]$ and RMSE $=$ $[0.84, 1.08]$. In terms of coefficient of determination, the median scores of our models are within those quoted ranges, though the RMSE tend to be higher. Note that Ma et al. (2021) used vegetation indices and weather variables as the predictive features and then trained the models using 10-years of data while the testing was performed on a dataset from a single year. In contrast, our models were trained using 2-years of data and tested for generalisability using 13-years of data.

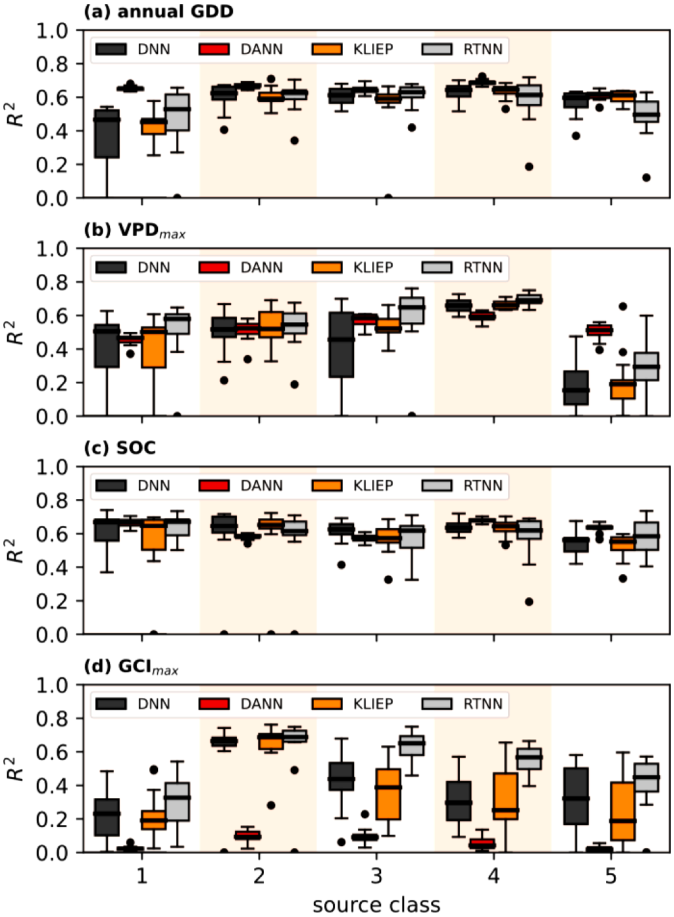


Fig. 6. Coefficient of determination (R^2) as a function of source classes categorised by (a) annual GDD, (b) maximum VPD, (c) SOC, and (d) maximum GCI.

3.2. *Inter-region generalisability*

More explainable variations of scores can be seen in Fig. 6 and Fig. 7 which summarize the results from the generalisability experiments using spatially segregated domains. Except for the bottom panel where the domains are categorised by the maximum GCI, the R^2 scores vary around 0.6, but with different spreads and trends. The typical value of RRMSE is above the one from the inter-annual generalisability experiments, that is $\gtrsim$15%. Apart from that, some important points can be extracted from the results.

Firstly, higher generalisability scores can be achieved by the models trained using the middle-class source (e.g. class 3) while the models trained using edge-classes tend to perform worse. However, the trends are not symmetrical. Models from regions with the lowest annual GDD (northernmost regions) have significantly lower scores than the models from the other end of classification. This also needs to be understood when trying to adapt machine learning models across regions with latitudinal difference.

Secondly, the generalisability scores slightly increase as a function of the maximum VPD class, but the scores for class-5 models drop significantly. If we refer to Fig. 1, class-1 VPD is associated with regions with the highest crop yields in the north-eastern part of US corn belt while the next classes are westward of those regions. The dataset extracted from each class has its own characteristics that determine the models' performance. Class-1 is associated with regions with relatively high yields with low statistical range/spread. The crop yields from class-2 to class-3

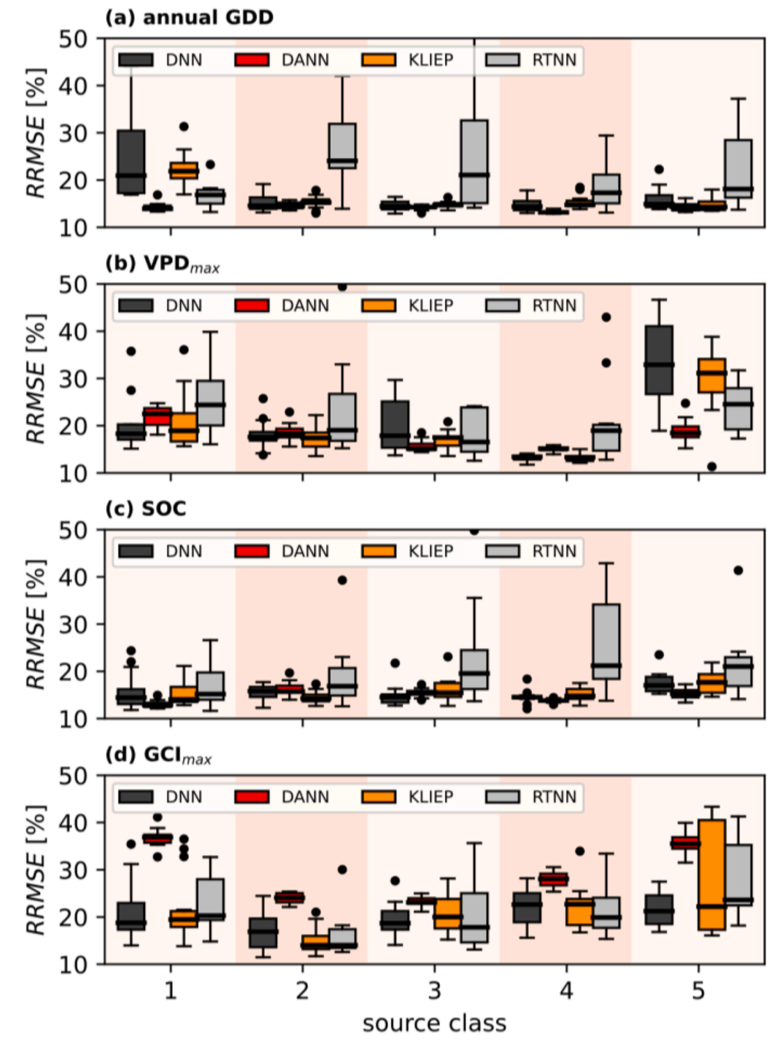


Fig. 7. Same as Fig. 6, but for RRMSE.

regions have a broader range such that the models have more chance to learn general patterns from the training datasets. On the other hand, class-5 occupies regions which are climatically categorised as great plains semi-arid eco-regions (Bolton and Friedl, 2013). The generalisability scores from the models trained in this class is significantly lower than others. This is in agreement with the results from Bolton and Friedl (2013) that correlations between vegetation indices (NDVI and EVI2) and crop yield drop significantly for the semi-arid regions. For such regions, NDWI serves as a better yield predictor. We can also compare this finding to the transferability experiments from Ma et al. (2021). They noted that model transfer from east regions (Eastern Temperate Forests, ETF) to the west (Great Plains, GP) tend to produce a lower coefficient of determination compared to the inverse transfer (GP → ETF).

Next, experiments using regions segregated by SOC produced generalisability scores which are more stable compared to the scores achieved in other categorisations. Slight decreases of R^2 and increases in RRMSE are observed for models trained using data from class 5. This class occupies more than half of the counties in North Dakota, Minnesota, and Wisconsin where the average crop yields are relatively low. Again, the lack of spread in crop yield distributions is associated with the lower performance scores, though the difference is small.

Lastly, the categorization based on the maximum GCI produced regions which are also distinguished in terms of crop yields. GCI and similar vegetation indices are known to have a good correlation with crop yield (Bolton and Friedl, 2013; Kang et al., 2020). This feature also highly correlates with other features used in this study, especially EVI, NDWI, VPD, LST, and Tmean, where the last three show negative correlation with GCI. Segregating regions according to GCI means dividing data into segments that rarely intercept with each other both in input and output space. Consequently, we cannot expect models with good generalisability scores if the models are trained using GCI-segregated datasets, even with a proper domain adaptation. As seen in Fig. 5, the median R^2 for DANN are close to zero while the scores for other models show wide spreads. These results emphasize that in order to obtain reliable models, we need to use training datasets which cover a wider range of input and output space. Domain adaptation has a capability to improve the models implemented on different targets by correcting domain shifts (Pan and Yang, 2009; Tuia et al., 2016). If the training datasets are too constrained, then the models face extrapolation problems where the prediction errors beyond the range of training datasets grow substantially.

As an additional test, we also fed the data from four selected states (Illinois, Indiana, Iowa, and Missouri) acquired in 2006–2015 into the models and tested the models using the data from all eleven states obtained in 2018–2019. The four states were selected considering the results from generalisability experiments, especially the fact that good models were usually trained using datasets from the temperate regions (not too cold, not too dry) with sufficient range of crop yields. Fig. 8 summarizes the comparison between the predicted and actual crop

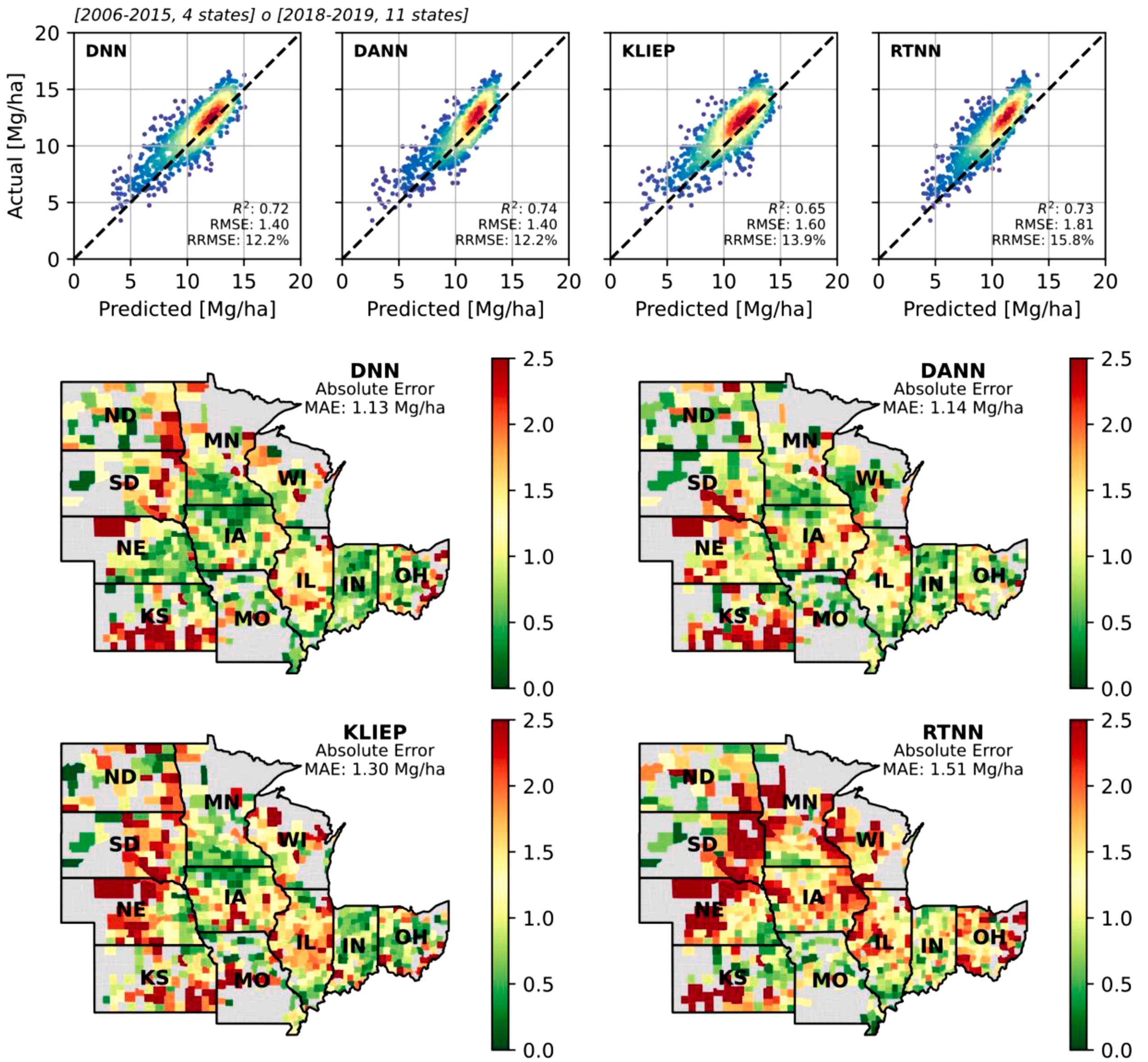


Fig. 8. Top: scatter plot of predicted versus actual crop yield in 2018–2019 with color coded density. The predictions are provided by four models (DNN, DANN, KLIEP, RTNN) trained using data from 4 states from 2006 to 2015. Below: map of the average absolute errors with the mean average error (MAE) printed on each panel.

yields. Small negative bias between the predicted and the actual crop yield can be observed. Most of the predicted values lay on the left side of the one-to-one dashed line in the scatter plot telling that the actual values are above the predicted ones. As seen in Fig. 1, the average crop yield in 2006–2014 is lower than the average yield in 2018–2019. This difference becomes the reason why the models predict lower yield than the actual values. Technological advancement or any improvement in managerial aspects may lead to a systematically higher crop yield that is not clearly reflected in remote sensing data we use. Except for KLIEP, the models achieved $R^2 > 0.70$. RTNN achieved $R^2 = 0.73$, but with RMSE $= 1.81$ Mg/ha (RRMSE $= 15.8\%$), which indicates more bias is introduced by this domain adapted model. Spatially, we can evaluate the model performances by calculating the absolute error ($|\hat{y}_i - y_i|$) and mapping the scores over the area of study. DNN and DANN are comparable though clear improvements (lower errors) can be observed in North Dakota (ND) and South Dakota (SD). Apart from that, KLIEP and RTNN show low accuracy for the western regions. In addition to the visual mapping, we also computed the Moran's I index (Moran, 1950; Anselin, 1995; Shermer, 2008) for spatial dependence of the observed errors. The results from all four models show clustering of the errors ($I_{\text{DNN}} = 0.39$, $I_{\text{DANN}} = 0.31$, $I_{\text{KLIEP}} = 0.32$, $I_{\text{RTNN}} = 0.41$) meaning that prediction errors correlate with geographic location. Effective domain adaptation like DANN shows a significant reduction of I with respect to the benchmark model (DNN).

3.3. *Transferability among diversity*

Domain shifts induce errors to crop yield predictions and domain adaptation algorithms try to suppress these errors. As in the classification tasks with domain adaptations, factors affecting the performances of machine learning models include the sufficiency (and quality) of labelled data for training and the degree of divergence between source and target datasets (Wang et al., 2019; Kluger et al., 2021). This issue was explored in our transferability experiments where pairs of source and target domains were generated and used for training and testing processes. Fig. 9 summarizes the scores achieved by four models as a function of Kullback-Leibler divergence (D_{KL}). The statistics (e.g. median scores) were calculated using scores achieved in the experiments using

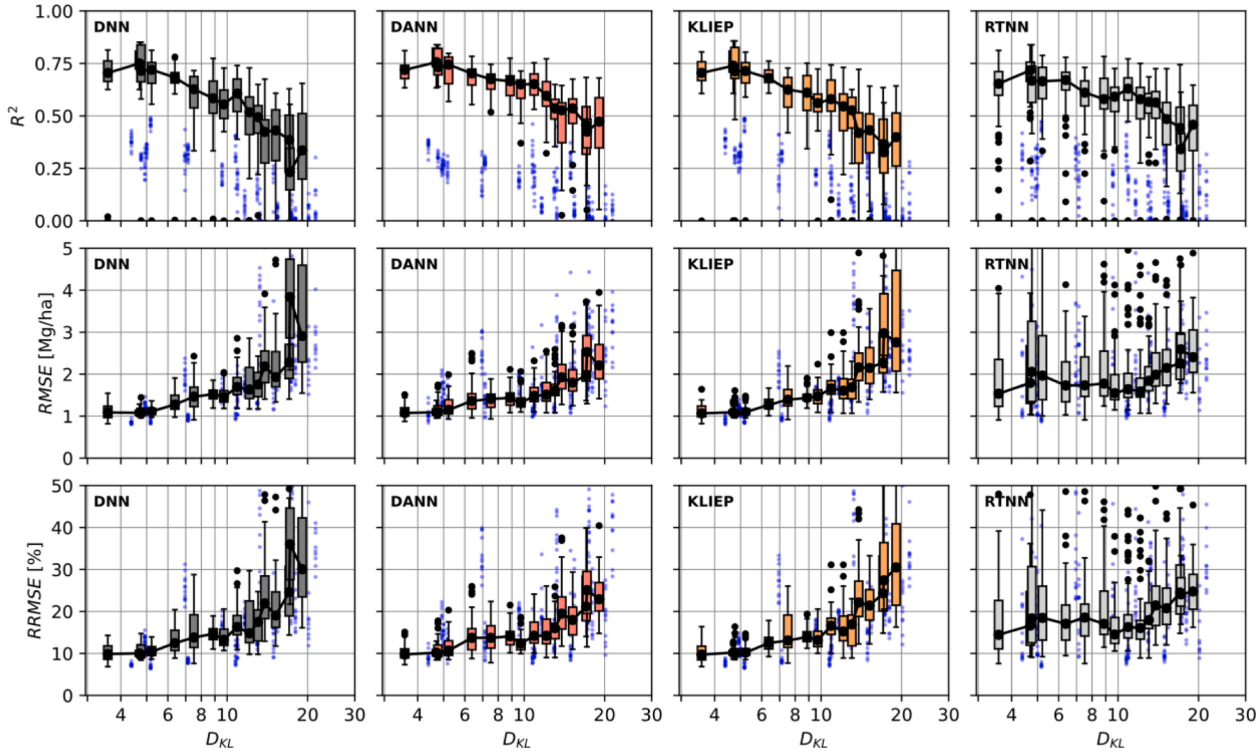


Fig. 9. Performance scores (above: R^2, middle: RMSE, bottom: RRMSE) as a function of Kullback-Leibler divergence between source and target domains. Blue dots represent the score from transferability experiments using GCI-segregated regions. Scores from experiments using GDD, VPD, and SOC as categorisers are binned and aggregated into boxplots with median values in very thick dots.

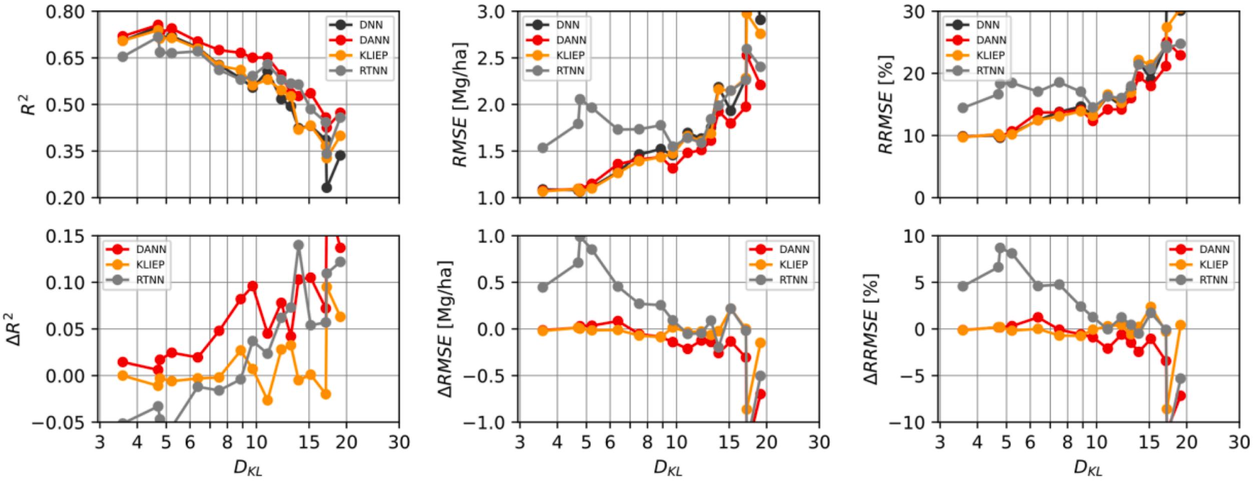


Fig. 10. Top panels: comparison between median scores (R^2, RMSE, RRMSE) achieved by four different models as a function of Kullback-Leibler divergence. Bottom panels: score differences relative to the baseline DNN.

regions categorised by the annual GDD, VPD, and SOC. Prior to the statistical aggregation, the results were binned into 16 equal-sized bins according to the divergence. Results from the experiments using temporal segregation were not included in the statistical aggregation since the domains in those experiments are less diverse. On the other hand, the categorization by GCI produces systematically lower scores as discussed before such that these results were not included in the statistical aggregation of performance scores in Fig. 9.

From Fig. 9, it is obvious that the performance deteriorates with increasing divergence between source and target domains, but domain adaptation algorithms generally perform their tasks by lowering the slope of decline. The spread of the scores also increases as indicated by the length of the boxplots. As emphasized by Ben-David et al. (2010), the target error is a combination of empirical source error and the error induced by the divergence. Direct comparison of the scores achieved by the models is presented in Fig. 10. To see how good are the improvements provided by the three domain adaptation algorithms with respect to the baseline model, the relative scores (ΔR^2, ΔRMSE, and ΔRRMSE) are also plotted. From this figure, we learn that the RTNN models tend to produce higher errors at low D_{KL}, mainly due to the random selection of labelled data from target domains. The bottom right panel of Fig. 9 also shows the RMSE with large variations produced by RTNN models. In other words, wrongly selected labelled data from the target domain may cause negative transfer where the score of the domain adapted model is lower than the traditional model. In terms of R^2, the domain adaptations slightly improve the performances at $D_{\mathrm{KL}} < 10$, while the improvements are more prominent at larger D_{KL}. Within the range of $10 \lesssim D_{\mathrm{KL}} \lesssim 20$, there is no sign of turning points of ΔR^2 and ΔRMSE implying that domain adaptations can be utilized effectively, possibly beyond this divergence range.

4. Discussion

4.1. Comparison to other studies

Results presented in Fig. 8 demonstrate how well the selected machine learning models predict maize yield in the US corn belt. With only ten features that capture the conditions in the peak season (August) as the input features, those models achieve performances that are comparable to the results in the literature. However, the performance comparisons are not always apple-to-apple as the performance score is highly dependent on the data used for both training and testing.

As a good representation from our study, the DANN acquires RMSE = 1,40 Mg/ha, RRMSE = 12.2%, $R^2 = 0.74$, and MAE = 1.14 Mg/ha. This model was trained using data from four states and tested using data from all eleven states acquired in different years. In recent literature, a well-trained machine learning model can achieve RMSE as low as 0.9 Mg/ha, especially when the training and testing datasets come from the same area but different time (Kang et al., 2020; Ansarifar et al., 2021; Ma et al., 2021). The common recipe to achieve this excellent performance includes the use of time-series data that capture the phenology of the crop, i.e. features extracted in different times starting from the vegetative to ripening stage. The inclusion of some managerial aspects can also be the key of success though these factors are rarely available. Time is a crucial factor in the crop yield prediction. A good model with 1 Mg/ha RMSE can deteriorate (RMSE > 1.5 Mg/ha) when it is laboured to provide prediction based on the features acquired at the vegetative stage (Kang et al., 2020; Ansarifar et al., 2021). In terms of RRMSE, the reported values range from 9% (Shahoseini et al., 2020) to 18% (Jin et al., 2017) depending on the spatial and temporal scope of the study. From a different model family, a crop model with data assimilation can achieve 1.24 Mg/ha RMSE and 11.5% RRMSE (Lu et al., 2022).

The above comparisons positioned the results of our current study in the pursuit of accurate and reliable predictions. Our results are not among the top, but they provide fairly good predictions only with ten features.

4.2. What to transfer

In this study, we demonstrated that domain adaptation improves the validity and performance of machine learning model for crop yield prediction, especially when the model is transferred to different target domain. Features that characterize soil (organic content, elevation), crop health (vegetation indices), and weather (temperature, humidity, etc.) conditions were used as the key inputs to the model. Most of those features are region-specific and highly dependent on the climate. Some are seasonally variable while the other can be assumed to be invariant for the whole year. We highlighted the importance of the ten features in predicting the maize yield in the US corn belt. GCI, EVI, NDWI, and FAPAR are features that capture the crop health. LST, GDD, and VPD represent the weather conditions that influence the yield. Lastly, SOC and Elevation are two features that characterize the soil. In the domain adaptation stage, those features can be transformed or re-weighted with/without the help of some crop yield data for training.

Except for the annual GDD, the features used in our machine learning models can be obtained immediately such that within-season crop yield prediction can be performed, assuming that crop type map is available (see Johnson and Mueller, 2021). For most of the corn crop in the US corn belt, the season peaks around August and the features obtained in August can be good proxies of the yield (Johnson, 2014; Kang et al., 2020). utilization of features obtained in August resulted in a sufficiently good crop yield prediction. However, we do not reject the notion that the peak vary by location, especially when a finer resolution (below county level) is addressed. Capturing the exact peak of season through phenology analysis is expected to slightly improve the prediction accuracy.

The urge of domain adaptation is not only raised by regional difference between source and target domains. Domain adaptation is required when we are dealing with different data sources or distinguished instruments for data acquisition. For instance, we used vegetation indices that are derived from MODIS datasets in this study while higher resolutions are widely available to capture the condition at crop-level (Kayad et al., 2019; Skankun et al., 2021). Transferring models that was trained using MODIS-based vegetation indices to work with a higher resolution remote sensing data can be a new challenge for domain adaptation, a challenge that needs to be addressed in future studies. The same necessity arises by the fact that weather parameters from PRISM dataset are only available for conterminous US. The use of different data source with global coverage, such as TerraClimate, is the only option available if we want to transfer the model to work outside the US without exhaustively re-training the model from scratch. Intuitively, the correlation between PRISM and TerraClimate data needs to be assessed prior to the transfer. A statistical model can be built to predict the PRISM-like weather parameters according to the available TerraClimate data in the target region. In this simple way, a model that was trained using PRISM data can be transferred to regions uncovered by PRISM.

More complex ways of transfer can also be constructed and evaluated. Recently, Ma et al. (2023) presented a case of multisource domain adaptation scheme to reduce the domain shift between US and Argentina in the context of maize yield. In that case, some weather parameters were acquired from different sources considering the dataset's geographical coverage. In the source domain where multisource data was available, two models were trained in parallel using two datasets from different sources. Adversarial training sequence was performed recursively to find predictive and domain-invariant features that are applicable in the target region.

4.3. How to transfer

Regarding the mechanism or the algorithm to transfer a machine learning model through domain adaptation, our results suggest that feature-based approach becomes the best alternative. On the other hand, instance-based approach with re-weighting process seems to be inferior. In an instance-based adaptation, there is a higher risk of overfitting as there are numerous possible samples of target data to be used for adjusting the model. Generally, domain adaptation using instance-based is less effective compared to the feature-based one (Pan and Yang, 2009; Bai et al., 2019).

Among the three domain adaptation algorithms invoked in this study, DANN gains the best scores and provides the most significant improvements compared to the baseline model (see Fig. 9 and Fig. 10). Statistically, this algorithm can reduce the RMSE by 0.7 Mg/ha relative to the DNN. RTNN is second best both in terms of R^2 and RMSE, except for the small divergence segment. If the algorithm is fed with more representative labelled data from the target domain, then the performance is expected to be higher. Lastly, KLIEP improves the coefficient of determination for $D_{\mathrm{KL}} \gtrsim 10$ while the improvement in terms of RMSE is less significant. Though these three algorithms cannot represent the whole domain adaptation, at least we can argue the following points about what and how to transfer the knowledge between domains. Firstly, a feature-based approach might be a good choice for domain adaptation since an additional learning process is imposed during feature encoding such that domain-invariant features can be revealed. Secondly, an instance-based approach may be an alternative, but its performance depends on how the distributions of source and target datasets intercept each other. The semi-supervised approach through

Table 2
List of states or provinces of some top corn-producing countries. The yields are national-averages from 2020/2021 statistics compiled by the Foreign Agriculture Service USDA. The provincial-level yields are higher than the national-average.

Country	States/Provinces	Maize Yield
China	Heilongjiang, Jilin, Liaoning	6.32 Mg/ha
Brazil	Mato Grosso	4.37 Mg/ha
India	Tamil Nadu, Karnataka	3.19 Mg/ha
South Africa	Free State, Mpumalanga, North West, Gauteng	5.42 Mg/ha
Nigeria	Kaduna, Niger, Katsina, Borno, Plateau	1.82 Mg/ha

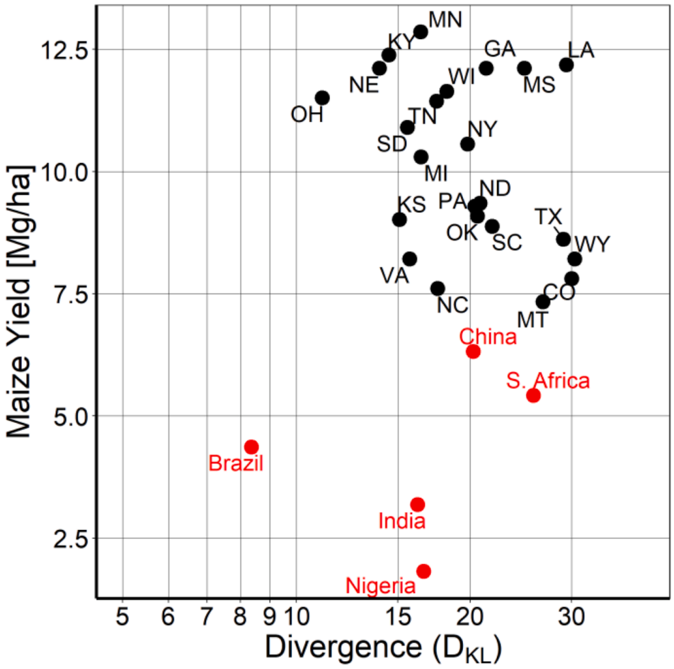


Fig. 11. Plot of 2020/2021 maize yields in some corn-producing states in the US (black) and some other countries (red) as a function of divergence relative to Illinois, Indiana, Iowa, and Missouri (as integral).

regularization may produce a good result if the model adapts to some well-labelled data from the target domain. Otherwise, regularization has a higher chance of producing a negative transfer model.

4.4. *When to transfer*

The realization of a predictive model which is generally applicable to many regions is perhaps the ultimate objective of data-driven modeling. Any machine learning model is expected to be transferable to different regions but we need to consider the difference between the regions in terms of relevant environmental and biophysical parameters.

Considering the fact that reliable crop yield records are not always available everywhere, transferable prediction models will be beneficial for agriculture monitoring and food security monitoring at the regional and global scales. To justify whether the model is transferable to a certain target domain, Kullback-Leibler divergence can be utilized. For the models discussed in this study, we can refer to Fig. 10 to justify the feasibility of transfer across domains. Suppose that a reliable model has performance scores of $R^2 \geq 0.5$ and relative RMSE of $\lesssim 20\%$ or approximately 2 Mg/ha, then we can expect that the baseline DNN can be transferred to the regions with $D_{\mathrm{KL}} \lesssim 12$ divergence relative to the source domain. Domain adapted models can be transferred further. DANN, for instance, can be transferred to regions with $D_{\mathrm{KL}} \lesssim 15$. To obtain a more geographically orientated picture of the Kullback-Leibler divergence, we computed the divergence between features extracted from selected regions in the US and the regions outside the study area. For this purpose, we extracted the same features from other corn producing states in the US and some other countries (see Table 2). Divergences between those regions relative to the four selected states in the US (Illinois, Indiana, Iowa, and Missouri) were calculated and then plotted in Fig. 11. The 2020 county level maize yields were acquired from NASS/USDA while the 2020/2021 national-level maize yields were acquired from the Foreign Agriculture Service USDA. It is worth noting that the selected states/provinces are among the top producers in each country such that the actual yields are higher than the national-average.

Relative to the selected four states in the US corn belt, Mato Grosso (Brazil) resembles them most closely with $D_{\mathrm{KL}} < 10$. The US mid-west states have D_{KL} between 10 and 20, while the mid-Atlantic states like New York (NY) and Pennsylvania (PA) have $D_{\mathrm{KL}} \gtrsim 20$. The north-eastern provinces (sheng) in China have similar climatic characteristics with the mid-Atlantic US such that the divergence between China and the reference regions is $\sim$20. Even though located at lower latitudes, states in India and Nigeria have intermediate divergence. The southern states like Georgia (GA), Mississippi (MS), and Louisiana (LA) have larger divergences compared to the central corn belt, while irrigated crops in these states have relatively high yields. The western states like Montana (MT), Wyoming (WY) and Colorado (CO) have D_{KL} close to 30 and produce less corn per acre. From this outlook, transferring machine learning models trained using US datasets to other countries has a reasonable basis though the divergence measure is not the only factor to determine the feasibility of transfer learning.

Alternatively, the divergence measures in general can also be utilized to select ideal source domains for training considering the characteristics of the target domain (Kluger et al., 2021). Suppose that we have several models trained using datasets from different source domains with various divergences relative to the target domain. In such a situation, we can select a model based on the least-divergence criteria and boost the performance through domain adaptation. For instance, a model trained using data from the western states of the US is expected to be a good stepping stone for domain adaptation to regions with similar climate characteristics such as South Africa.

4.5. *Limitations*

This study demonstrates how domain adaptation can improve the performance of machine learning-based crop yield prediction models. This approach opens the wider possibility for transferring a well-trained model from a data rich region to different regions. However, there are limitations of this study which can be improved further.

Firstly, we utilised ten features in the machine learning models. Among those ten features, there are dynamic features like vegetation indices and weather parameters that change over time. However, we only use the maximum values of those dynamic features in the model. Even though the features were selected carefully according to their importance, additional features may be beneficial to improve the models when applied to other regions. For example, Anghileri et al. (2022) showed the significance of precipitation as the predictor of crop yield in Malawi, Africa. Other hydrological parameters such as soil moisture also shows a higher correlation to crop yield in Zambia (Vergopolan et al., 2021). In this study, VPD was selected as the drought sensitive variable (Lobell et al., 2014). Though it has a moderate anti-correlation with precipitation and soil moisture, inclusion of more weather and hydrological variables is expected to produce more generalisable models.

Secondly, the use of MODIS data to derive vegetation indices is considered to be sufficient for the case of county level maize yield in the US corn belt. In this region, large scale crops are common such that 250-m resolution images are adequate to capture the crop health at any time. However, the utilization of higher resolution images may be necessary when we are dealing with small scale cropping practices as commonly found in developing regions (Jin et al., 2019). The accurate

Table A1
Complete list of features analysed in this study, including the ones not selected as the predictors. Feature importance based on the correlation to the crop yield (R^2) and the mean decrease in accuracy (MDA) from fitting machine learning model are also displayed.

Feature	R^2	MDA	Formula	Sources
Green Chlorophyll Index (GCI)	0.76	0.67	$\frac{B_2}{B_4}-1$	MODIS Nadir Bidirectional Reflectance Distribution Function Adjusted Reflectance (MCD43A4.006) Resolution: 500 m, daily Source: 10.5067/MODIS/MCD43A4.006 B1: red (620–670) B2: NIR (841–876 nm) B3: blue (459–479 nm) B4: green (545–565 nm) B6: SWIR (1628–1652 nm)
Enhanced Vegetation Index (EVI)	0.68	0.04	$\frac{2.5(B_2-B_1)}{(B_2+6B_1-7.5B_3+1)}$	
Normalized Difference Water Index (NDWI)	0.74	0.45	$\frac{B_2-B_6}{B_2+B_6}$	
Normalized Difference Vegetation Index (NDVI)	0.63	0.02	$\frac{B_2-B_1}{B_2+B_1}$	
Leaf Area Index (LAI)	0.33	0.01		MODIS Leaf Area Index product (MOD15A2H). Resolution: 500 m, 8 days Source: 10.5067/MODIS/MOD15A2H.006
Fraction of Photosynthetically Active Radiation (FAPAR)	0.51	0.01		
Daytime Land Surface Temperature (LST)	0.41	0.01		MODIS Terra Land Surface Temperature and Emissivity (MOD11A1.006) Resolution: 1000 m, daily Source: 10.5067/MODIS/MOD11A1.006
Maximum Temperature (Tmax)	0.23	0.01		Parameter-elevation Relationships on Independent Slopes Model (PRISM, Daly et al., 2008, 2015) Resolution: 4000 m, daily
Average Temperature (Tmean)	0.14	0.03		
Growing Degree Date (GDD)	0.06	0.01	$\sum \min(\max(Tmean,10),30)-10$	
Vapor Pressure Deficit (VPD)	0.30	0.03		
Bulk density of the fine earth fraction (BDOD)	0.02	0.01		SoilGrids (Hengl et al., 2017) Resolution: 250 m, single epoch: 2000
Volumetric fraction of coarse fragments (CFVO)	0.02	0.01		
Proportion of clay particles (CLAY)	0.01	0.01		
Total nitrogen (NITRO)	0.01	0.02		
Proportion of sand particles (SAND)	0.01	0.01		
Proportion of silt particles (SILT)	0.01	0.01		
Soil Organic Content (SOC)	0.01	0.02		
Elevation	0.01	0.05		SRTM digital elevation model Resolution: 90 m, single epoch: 2000 Source: https://srtm.csi.cgiar.org

representation of vegetation indices from the crop is also reliant on the cropland identification and masking. For the case of the US corn belt, CDL becomes the best choice for that purpose though this map is available months after the harvest. Consequently, an independent crop map is required for mid-season maize yield prediction (Schwalbert et al., 2020). The same applies for the extraction of features from regions outside the US where the global cropland mask is not sufficient.

Lastly, this study utilizes data from the US corn belt which is associated with high yield agricultural practices. As reviewed by Lobell et al. (2009), the maize yield in this region is more than 40% of the potential yield derived using an idealized crop model. When transferring the model to regions with significantly higher yield gaps, additional adjustment or scaling will be required. From the perspective of crop modeling, the crop yield is regarded as a fraction of above-ground biomass that is harvested. The biomass itself is a product of the green canopy development and the crop transpiration (e.g., Vanuytrecht et al., 2014). These variables can be estimated from remote sensing data while the final conversion of biomass to crop yield requires a harvest index which depends on many factors including cultivars.

5. Conclusion

Data-driven models, especially machine learning models have become an important part of modern agriculture monitoring and forecasting which aim to help achieve food security and sustainability. Beyond conventional modeling, domain adaptation scheme may increase the generalisability and transferability of the models.

In this study, we conducted systematic generalisability and transferability experiments using data from US corn belt segregated by features namely year, annual grow degree days (GDD), vapor pressure deficit (VPD), soil organic content (SOC), and the green chlorophyll vegetation index (GCI). We trained traditional deep neural network (DNN) as the baseline, together with neural network equipped with three domain adaptation algorithms, namely Discriminative Adversarial Neural Network (DANN), Kullback-Leibler Importance Estimation Procedure (KLIEP), and Regular Transfer Neural Network (RTNN). Those models were trained using data from a specific source domain and trained to a different target domain where the difference between those domains was measured using Kullback-Leibler divergence. We found that the models trained using data from recent years (since 2014) tend to have lower generalisability scores. Unaccounted factors such as variation in precipitation, field managements and intensification may contribute to the variation of maize yields. We identified that models trained using data from colder regions with lower annual GDD lack generalisability. The same applies for models trained using data from dry semi-arid regions where the VPD is higher. The spread of training data in input and output space is also a determining factor of model performance as indicated by the results from experiments using GCI-segregated regions. Training using a subset of data with a limited range of GCI produces models with low performance.

In general, the performance of machine learning models deteriorates as the divergence between source and target domains increases. Domain adaptations can alleviate this problem. Among three approaches evaluated in this study, re-weighting is the least preferable approach because it does not significantly improve the transferability. On the other hand, a feature-based approach provided the best performance. Parameter-based domain adaptation as a semi-supervised approach has fluctuating performance due to random selection of labelled data from the target domain. Consequently, this approach requires carefully selected training data to gain comparable performance. The domain adaptation approach is expected to improve transferability of a machine learning model trained in the US (or other regions with sufficient data) to data poor regions of the world to estimate crop yield.

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.

Acknowledgement

This work was funded through the “A new paradigm in precision agriculture assimilation of ultra-fine resolution data into a crop-yield forecasting model" project, supported by the King Abdullah University of Science and Technology, grant number OSR-2017-CRG6, and through the “Building REsearch Capacity for sustainable water and food security In drylands of sub-Saharan Africa” (BRECcIA) project which is supported by UK Research and Innovation as part of the Global Challenges Research Fund, grant number NE/P021093/1.

Appendix

Table A1

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33 (1), 121–131.

Anghileri, D., Bozzini, V., Molnar, P., Sheffield, J., 2022. Comparison of hydrological and vegetation remote sensing datasets as proxies for rainfed maize yield in Malawi. Agric. Water Manage. 262, 107375. March 2021.

Ansarifar, J., Wang, L., Archontoulis, S.V., 2021. An interaction regression model for crop yield prediction. Sci. Rep. 11 (1), 17754.

Anselin, L., 1995. Local indicators of spatial association-LISA. Geogr. Anal. 27, 93–115.

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141.

Bai, J., Cao, R., Ma, W., Shinnou, H., 2019. Combination of Feature-based and Instance-based methods for Domain Adaptation in Sentiment Classification. In: Proceedings of the 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, pp. 1–4 t.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114 (6), 1312–1323.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79 (1–2), 151–175.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84.

Boryan, C., Yang, Z., Mueller, R., Craig, M, 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26 (5), 341–358.

Chelba, C., Acero, A., 2006. Adaptation of maximum entropy capitalizer: little data can help a lot. Comput. Speech Language 20 (4), 382–399.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151 (June), 61–69.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. Int. J. Climatol. 28 (15), 2031–2064.

Daly, C., Smith, J.I., Olson, K.V., 2015. Mapping atmospheric moisture climatologies across the conterminous united states. PLoS One 10 (10), e0141140.

de Mathelin, A., Deheeger, F., Richard, G., Mougeot, M., and Vayatis, N. (2021). Adapt: awesome domain adaptation python toolbox. arXiv preprint arXiv:2107.03049.

Duncan, J., Dash, J., Atkinson, P.M., 2015. The potential of satellite-observed crop phenology to enhance yield gap assessments in smallholder landscapes. Front. Environ. Sci. 3, 56.

EPA. (2001). United states environmental protection agency. Quality Assurance Guidance Document-Model Quality Assurance Project Plan for the PM Ambient Air, 2.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larocehlle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Machine Learn. Res. 17, 1–35.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 315–323. JMLR Workshop and Conference Proceedings.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Attribute selection impact on linear and nonlinear regression models for crop yield prediction. Scientific World J. 2014.

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.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al., 2017. Soilgrids250m: global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.

Hunter, M.C., Smith, R.G., Schipanski, M.E., Atwood, L.W., Mortensen, D.A., 2017. Agriculture in 2050: recalibrating targets for sustainable intensification. Bioscience 67 (4), 386–391.

Izquierdo-Verdiguier, E., Laparra, V., Gomez-Chova, L., Camps-Valls, G., 2013. Encoding invariances in remote sensing image classification with SVM. IEEE Geosci. Remote Sens. Lett. 10 (5), 981–985.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., Kim, S.H., 2016. Random forests for global and regional crop yield predictions. PLoS One 11 (6), 1–15.

Jin, Z., Azzari, G., Lobell, D.B., 2017. Improving the accuracy of satellite-based high-resolution yield estimation: a test of multiple scalable approaches. Agric. For. Meteorol. 247, 207–220.

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google earth engine. Remote Sens. Environ. 228, 115–128. September 2018.

Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128.

Johnson, D.M., 2016. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. Int. J. Appl. Earth Obs. Geoinf. 52, 65–81.

Johnson, D.M., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218-219, 74–84.

Johnson, D.M., Mueller, R., 2021. Pre-and within-season crop type classification trained with archival land cover information. Remote Sens. Environ. 264, 112576.

Ju, S., Lim, H., Ma, J.W., Kim, S., Lee, K., Zhao, S., Heo, J., 2021. Optimal county-level crop yield prediction using MODIS-based variables and weather data: a comparative study on machine learning models. Agric. For. Meteorol. 307 (May), 108530.

Kang, Y., Khan, S., Ma, X., 2009. Climate change impacts on crop yield, crop water productivity and food security – a review. Prog. Nat. Sci. 19 (12), 1665–1674.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., Anderson, M, 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environ. Res. Lett. 15 (6), 064005.

Kayad, A., Sozzi, M., Gatto, S., Marinello, F., Pirotti, F., 2019. Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. Remote Sens. 11 (23), 2873.

Kingma, D.P. and Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kluger, D.M., Wang, S., Lobell, D.B., 2021. Two shifts for crop mapping: leveraging aggregate crop statistics to improve satellite-based maps in new regions. Remote Sens. Environ. 262 (May), 112488.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.

Kumar, M., 2016. Impact of climate change on crop yield and role of model for achieving food security. Environ. Monit. Assess. 188 (8), 465.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, W., Zhang, X., Peng, Y., Dong, M, 2021. Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms. Int. J. Remote Sens. 42 (6), 1973–1993.

Lobell, D.B., Cassman, K.G., Field, C.B., 2009. Crop yield gaps: their importance, magnitudes, and causes. Annu. Rev. Environ. Resour. 34, 179–204.

Lobell, D.B., Roberts, M.J., Schlenker, W., Braun, N., Little, B.B., Rejesus, R.M., Hammer, G.L., 2014. Greater sensitivity to drought accompanies maize yield increase in the US. Midwest. Science 344 (6183), 516–519.

Ma, Y., Zhang, Z., Yang, H.L., Yang, Z, 2021. An adaptive adversarial domain adaptation approach for corn yield prediction. Comput. Electron. Agric. 187 (May), 106314.

Lu, Y., Wei, C., McCabe, M.F., Sheffield, J., 2022. Multi-variable assimilation into a modified AquaCrop model for improved maize simulation without management or crop phenology information. Agricultural Water Management 266, 107576.

Ma, Y., Yang, Z., Zhang, Z., 2023. Multisource maximum predictor discrepancy for unsupervised domain adaptation on corn yield prediction. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Meroni, M., Waldner, F., Seguini, L., Kerdiles, H., Rembold, F., 2021. Yield forecasting with machine learning and small data: what gains for grains? Agric. For. Meteorol. 308–309 (April).

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23.

Mueller, N.D., Gerber, J.S., Johnston, M., Ray, D.K., Ramankutty, N., Foley, J.A., 2012. Closing yield gaps through nutrient and water management. Nature 490 (7419), 254–257.

Pacifici, F., Longbotham, N., Emery, W.J., 2014. The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images. IEEE Trans. Geosci. Remote Sens. 52 (10), 6241–6256.

Pan, S.J.P., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Park, S., Im, J., Park, S., Yoo, C., Han, H., Rhee, J., 2018. Classification and mapping of paddy rice by combining Landsat and SAR time series data. Remote Sens. (Basel) 10 (3), 1–22.

Perez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. IEEE International Symposium on Information Theory - Proceedings, pages 1666–1670.

Qader, S.H., Dash, J., Atkinson, P.M., 2018. Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: a case study in Iraq. Sci. Total Environ. 613-614, 250–262.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566 (7743), 195–204.

Schwalbert, R., Amado, T., Nieto, L., Corassa, G., Rice, C., Peralta, N., Schauberger, B., Gornott, C., Ciampitti, I., 2020. Mid-season county-level corn yield forecast for us corn belt integrating satellite imagery and weather variables. Crop Sci. 60 (2), 739–750.

Shahhosseini, M., Hu, G., Archontoulis, S.V., 2020. Forecasting corn yield with machine learning ensembles. Front. Plant Sci. 11, 1120.

Shermer, M., 2008. Patternicity: finding meaningful patterns in meaningless noise. Sci. Am. 299 (5), 48.

Skakun, S., Kalecinski, N.I., Brown, M.G., Johnson, D.M., Vermote, E.F., Roger, J.C., Franch, B., 2021. Assessing within-field corn and soybean yield variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 satellite imagery. Remote Sens. (Basel) 13 (5), 872.

Sugiyama, M., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. Proceedings of the 20th International Conference on Neural Information Processing Systems, pages 1433–1440.

Teluguntla, P., Thenkabail, P.S., Xiong, J., Gumma, M.K., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T.T., et al., 2015. Global Cropland Area Database (GCAD) Derived from Remote Sensing in Support of Food Security in the Twenty-First century: Current Achievements and Future Possibilities. Taylor & Francis.

Teng, W., Wang, N., Shi, H., Liu, Y., Wang, J, 2020. Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images. IEEE Geosci. Remote Sens. Lett. 17 (5), 789–793.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: an overview of recent advances. IEEE Geosci. Remote Sens. Magazine 4 (2), 41–57.

UN, 2019. World Population Prospects 2019: Highlights, 11. United Nations Department for Economic and Social Affairs, New York (US), p. 125.

USDA, 2022. Crop production historical track records. National Agricultural Statistics Service, United States Department of Agriculture, Washington, DC.

van Dijk, M., Morley, T., Rau, M.L., Saghai, Y., 2021. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nat Food 2 (7), 494–501.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. 177 (August), 105709.

Vanuytrecht, E., Raes, D., Steduto, P., Hsiao, T.C., Fereres, E., Heng, L.K., Vila, M.G., Moreno, P.M., 2014. Aquacrop: fao's crop water productivity and yield response model. Environ. Model. Softw. 62, 351–360.

Vergopolan, N., Xiong, S., Estes, L., Wanders, N., Chaney, N.W., Wood, E.F., Konar, M., Caylor, K., Beck, H.E., Gatti, N., Evans, T., Sheffield, J., 2021. Fieldscale soil moisture bridges the spatial-scale gap between drought monitoring and agricultural yields. Hydrol. Earth Syst. Sci. 25 (4), 1827–1847.

Walker, J., De Beurs, K., Wynne, R., Gao, F., 2012. Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology. Remote Sens. Environ. 117, 381–393.

Wang, A.X., Tran, C., Desai, N., Lobell, D., and Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, pages 1–5.

Wang, S., Azzari, G., Lobell, D.B., 2019. Crop type mapping without field-level labels: random forest transfer and unsupervised clustering techniques. Remote Sens. Environ. 222, 303–317. December 2018.

Xu, H., Zhang, X., Ye, Z., Jiang, L., Qiu, X., Tian, Y., Zhu, Y., Cao, W., 2021. Machine learning approaches can reduce environmental data requirements for regional yield potential simulation. Eur. J. Agron. 129. August 2020.

Yang, W., Nigon, T., Hao, Z., Dias Paiao, G., Fernández, F.G., Mulla, D., Yang, C, 2021. Estimation of corn yield based on hyperspectral imagery and convolutional neural network. Comput. Electron. Agric. 184. February.

Yoosefzadeh-Najafabadi, M., Earl, H.J., Tulpan, D., Sulik, J., Eskandari, M., 2021. Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. Front. Plant Sci. 11 (January), 1–14.

You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). Deep gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the Thirty-First AAAI conference on artificial intelligence.

Zhao, Y., Potgieter, A.B., Zhang, M., Wu, B., Hammer, G.L., 2020. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. Remote Sens (Basel) 12 (6), 1024.