County-Level Corn Yield Prediction with Deep Learning

By

Yuchi Ma

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Biological Systems Engineering)

at the

University of Wisconsin-Madison

2023

Date of final oral examination: 02/28/2023

The dissertation is approved by the following members of the Final Oral Committee:

Zhou Zhang, Assistant Professor, Biological Systems Engineering

Qunying Huang, Associate Professor, Geography

Jingyi Huang, Assistant Professor, Soil Science

Zhengwei Yang, USDA National Agricultural Statistics Service

Contents

To my parents, Meining Wang and Jun Ma, for their unconditional love

ACKNOWLEDGMENTS

The road toward a Ph.D. is long and winding. I am very lucky and glad to thank those who have accompanied me during this journey. First and foremost, words cannot express my gratitude to my advisor and chair of my committee Dr. Zhou Zhang for her invaluable help and support during the last four years. I am very lucky to join her group as her first Ph.D. student. She has provided me with an inclusive and professional research environment where I have been trained to become an independent researcher. Her invaluable scientific insights have inspired me a lot for addressing the grand challenges our planet faces.

I also could not have undertaken this journey without my committee members, Dr. Qunying Huang, Dr. Jingyi Huang, and Dr. Zhengwei Yang, who generously provided knowledge and expertise. Dr. Qunying Huang opens a new window for me to learn geospatial databases and cloud computing. It is my honor to serve as a lecturer for her class, which is my first teaching experience. Dr. Jingyi Huang's expertise in soil science has provided me with insightful comments on how to use soil properties and how to explain the driving factors during model development. I am very lucky to cooperate with him on some proposals. Dr. Zhengwei Yang, as a senior scientist from the USDA, has always inspired me to think critically about my research. He always asks me deep questions and provides perspicacious comments which help improve my work to a new level. I am very grateful for having the privilege to closely work with these great scientists and learn perspectives, ideas, and knowledge from them.

Moreover, I've been fortunate to work with many incredible people, Jing Zhou, Jiahao Fan, Yijia Xu, Tong Yu, Yu Li, and Parker Williams from BSE as well as visiting scholars Luwei Feng, Yumiao Wang, and Chen Sun. We have a lot of memorable field trips and research

discussions. It’s also been such a joy to have amazing friends who always support me, Bo Peng, Jinmeng Rao, Yuhao Kang, Yunlei Liang, Uling Tang, Jie Hu, and Yan Li, during grad school at UW-Madison.

Lastly, I would like to express my heartfelt gratitude to my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process. Your encouragement, understanding, and belief in me have been a source of strength and motivation. I would give my spatial thanks to my significant half for her love, encouragement, and willingness to share every moment of my life whether sad or joyful. Your presence in my life has made it all the more meaningful. I am so grateful to have you by my side. Finally, I would like to take a moment to honor and thank my grandfather, who passed away at the end of 2022. Despite his physical absence, he continues to be a source of strength and inspiration in my life. I will always be grateful for the time we shared and the memories that will live on. This achievement is dedicated to his memory. He always told me that he missed me and frequently asked me when I would get back home. Hi, grandpa, I miss you too and I will see you again.

The things that I have obtained from completing the dissertation are much more than earning a Ph.D. degree. They are the treasure of my life and will stay with me for the rest.

ABSTRACT

As the world's leading corn producer, the United States supplies more than 30% of the global corn production. Accurate and timely estimation of corn yield is therefore essential for commodity trading and global food security. Recently, machine learning (ML) and deep learning (DL) models have been explored for corn yield prediction. Despite the success, there are still two major limitations of existing ML-based crop yield prediction models. First, most existing models mainly focus on predicting the crop yield without providing any information about the uncertainty which is important to provide quantified confidence interval of the prediction to users for their knowledgeable decision makings. Second, data-driven DL models require a large amount of reference data samples (i.e., yield records) for model training and tend to have low spatial transferability due to domain shifts between different regions. In this dissertation, we focused on addressing these two major limitations using Bayesian learning and unsupervised domain adaptation (UDA) for corn yield prediction.

Specifically, to address the first limitation, this dissertation proposed Bayesian neural networks (BNN) for corn yield prediction and uncertainty analysis. By applying Bayesian inference, the proposed BNN model can provide not only accurate yield prediction but also the corresponding predictive uncertainty. Feature variables were collected from multiple data sources, including remote sensing (RS) imagery, weather variables, soil properties, and historical average yield. Using preceding years since 2001 for model training, the developed BNN model achieved an average coefficient of determination (R^2) of 0.77 for late-season prediction across the U.S. corn belt in testing years 2010–2019 and outperformed five other state-of-the-art ML models. Evaluation results of in-season yield prediction showed that the BNN model achieved the optimal prediction results by the middle of August, which is about two months before the

harvest. We also assessed the predictive uncertainty and found that more than 84% of the observed yield records were successfully enveloped in the 95% confidence interval of the predictive yield distribution. Uncertainties in yield prediction were mainly induced by the observation noise and related to the inter-annual and seasonal variabilities of environmental stress such as heat stress and water stress.

To address the second limitation, this dissertation utilized the UDA strategy to reduce the domain shift between the source domain and the target domain with the aim of accurately predicting corn yield in the target domain without using labeled data from the target domain. We first proposed two single-source UDA models for county-level corn yield prediction based on RS images and weather variables. The proposed adaptive domain adversarial neural network (ADANN) and Bayesian domain adversarial neural network (BDANN) have been proven to have better spatial transferability and outperformed other supervised learning models and DANN in transfer experiments across two ecoregions in the U.S. corn belt. Furthermore, we proposed a multi-source UDA method named multi-source maximum predictor discrepancy (MMPD) to address the remaining issues of single-source domain adaptation methods. First, the multi-source UDA strategy was adopted in MMPD to avoid negative interference among source samples from heterogeneous regions. Also, by using the maximum predictor discrepancy (MPD), MMPD was trained to align source and target domains by considering crop yield response in the target domain based on task-specific regression models. Experiments on three UDA scenarios in the U.S. corn belt and Argentina have been conducted to evaluate the model performance. It was observed that MMPD outperformed representative single-source and multi-source UDA methods.

In summary, this dissertation introduced Bayesian inference and UDA to county-level corn yield prediction based on RS and weather variables. Novel solutions have been provided for quantifying predictive uncertainty in crop yield prediction and improving spatial transferability for deep learning-based crop yield prediction models. This dissertation provides a robust framework for the in-season prediction of crop yield and highlights the need for a deeper understanding of the impact of environmental stress on agricultural productivity and crop yield. Moreover, this dissertation applied the UDA for crop yield prediction and demonstrated the effectiveness of adversarial learning for improving the transferability of DL models on crop yield prediction.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADANN	Adaptive Domain Adversarial Neural Network
AWC	Soil Available Water Holding Capacity
BatchNorm	Batch Normalization Layer
BDANN	Bayesian Domain Adversarial Neural Network
BNN	Bayesian Neural Network
CDL	Cropland Data Layer
CEC	Cation Exchange Capacity
CHIRPS	Climate Hazards Group Infrared Precipitation with Stations
CNN	Convolutional Neural Networks
DANN	Domain Adversarial Neural Network
DL	Deep Learning
DNN	Deep Neural Networks
EPA	Environmental Protection Agency
ET	Daily Mean Evapotranspiration
ETF	Eastern Temperate Forests
EVI	Enhanced Vegetation Indices
GEE	Google Earth Engine
GCI	Green Chlorophyll Index
GLDAS	Global Land Data Assimilation System
GLDASws	GLDAS Water Stress
GP	Great Plains
LSTday	Daytime Land Surface Temperature
LSTnight	Nighttime Land Surface Temperature
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MARE	Mean Absolute Relative Error
MCD	Maximum Classifier Discrepancy
MIR	Multiple Instance Regression
MMPD	Multi-source Maximum Predictor Discrepancy
MODIS	Moderate Resolution Imaging Spectroradiometer
MPD	Maximum Predictor Discrepancy
M3SDA	Moment Matching for Multi-Source Domain Adaptation
ML	Machin Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NASS	National Agricultural Statistics Service
NEON	National Ecological Observatory Network
NDVI	Normalized Difference Vegetation Indices

NDWI	Normalized Difference Water Indices
NIR	Near Infrared
NN	Neural Network
PDA	Partial Domain Adaptation
PPT	Daily Total Precipitation
PRISM	Parameter elevation Regressions on Independent Slopes Model
R^2	Coefficient of Determination
RA	Research Activity
RQ	Research Question
ReLU	Rectified Linear Unit
RF	Random Forests
RMSE	Root Mean Square Error
RS	Remote Sensing
SMPD	Single-source Maximum Predictor Discrepancy
SOM	Soil Organic Matter
SSURGO	Soil Survey Geographic database
SVR	Support Vector Regression
SWIR	Shortwave Infrared
TL	Transfer Learning
Tmean	Mean Air Temperature
Tmax	Maximum Air Temperature
t-SNE	t-distributed Stochastic Embedding
UDA	Unsupervised Domain Adaptation
USDA	United States Department of Agriculture
VPD	Vapor Pressure Deficit
VI	Vegetation Indices
WDRVI	Wide Dynamic Ranged Vegetation Index
XGBoost	Extreme Gradient Boosting

CHAPTER 1 INTRODUCTION

1.1 Crop Yield Prediction

With an increasing world population, it is projected that the world needs to feed 9.0 billion people by 2050 (Godfray, et al., 2010). Ending hunger and improving food security are among the prime sustainable development goals of the United Nations (Lu et al., 2015). As the world's largest corn producer and exporter, the U.S. harvested 366.6 million metric tons of corn and supplied over 30% of global corn production (Li et al., 2019; USDA, 2020a). Accurate and timely estimation of corn yield in the U.S. is therefore of great importance for farming resource management, food security monitoring, and market planning (Jiang et al., 2019; Johnson, 2014). Specifically, the accurate estimate of yields allows for better understanding of the food supply which in turn helps the demand side plan to better utilize the finite crop resources (Liu et al., 2021). Moreover, under the pressures of global warming and climate extremes (Crane-Droesch, 2018; Lobell et al., 2013, 2009), corn production in the U.S. has experienced substantial loss with increasing inter-year variability (Lobell et al., 2014; Sibley et al., 2014; USDA, 2020a). Seasonal estimation of large-scale corn yield can facilitate better assessments of its response to environmental stresses (Guan et al., 2017), and thus provide reliable information for adaptations in cropping systems for sustainable agriculture (Kang et al., 2020; Wang et al., 2018). Overall, accurate predictions of crop yield for corn and soybean in the U.S. are critical for ensuring food security, economic stability, and sustainable agricultural practices in the country.

There are several publicly available corn yield predictions in the U.S. For example, the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) publishes crop progress and yield predictions of major staple crops for the U.S. at

monthly schedules before harvest. For example, corn progress reports are published from May to November. However, these publicly available predictions are mainly at the national and/or at the state level, which cannot meet the need for precision agriculture and knowledgeable decision makings at local (county) level. Also, USDA crop yield prediction is derived based on the nationwide agricultural survey and monthly field surveys during the growing season, which are costly and labor-intensive. Moreover, USDA NASS reports annual county-level yield statistics (Figure 1-1), which is based on a large-scale survey that is conducted after the growing season. The annual county-level yield statistics report is not available to the public until February of the next year. As a result, the applications and decision-makers that require near real-time in-season yield predictions (e.g., to assess the impact of ad hoc disaster events on crop yield loss), become very difficult.

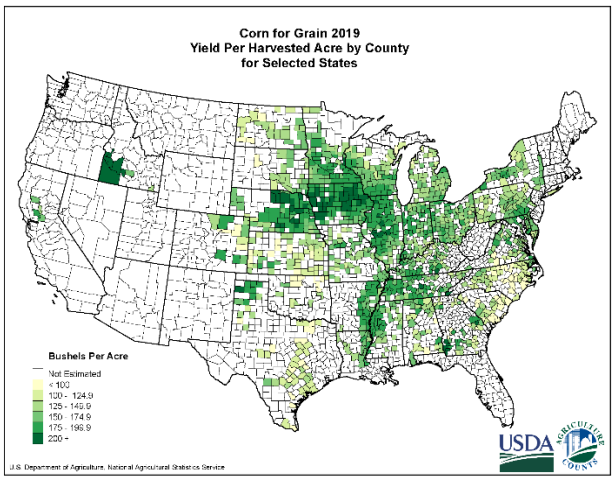

Figure 1-1. Corn yield in 2019 by county published by USDA NASS.

To provide in-season corn yield prediction, many yield prediction methods have been developed. They can be mainly categorized into two types, including physical simulation models and statistical machine learning (ML) models (Archontoulis et al., 2020; Feng et al., 2020; Maimaitijiang et al., 2020). Physical crop models are developed according to the physiological characteristics of crops and estimate yield by simulating the underlying crop and environmental

processes, such as crop growth, nutrient cycling, soil-plant dynamics, and water balance (Archontoulis et al., 2020; Liu et al., 2022; Zhang et al., 2019). Representative crop models include the CROPGRO-soybean model (Jagtap and Jones, 2002), CERES-Maize model (Hodges et al., 1987), and DSSAT-CROPGRO-Perennial Forage model (Malik et al., 2018). Although these models can well explain and estimate crop productivity based on biophysical processes, extensive locally sensed data related to both biotic and abiotic factors are required for model calibration (Cai et al., 2017; Kang and Özdoğan, 2019), limiting their applicability in large-scale yield modeling (Sakamoto et al., 2013; Zhang et al., 2019).

Statistical ML models, instead of simulating biophysical processes, attempt to perform yield estimation by establishing empirical relationships between driving factors of crop yield with historical yield records (Sun et al., 2020; Zhou et al., 2022). Therefore, they have the advantage of predicting crop yield with no need for explicit programming or knowledge of physiological mechanisms on individual crops (Wang et al., 2020; Zhang et al., 2019). Also, the advancement of satellite remote sensing (RS) technologies has enabled large-scale agricultural land monitoring with high spatial and temporal resolutions (Chen et al., 2022; Schwalbert et al., 2020). RS of crop canopies offers insights into plant canopy structure and crop health, as variations in canopy density and discoloration associated with plant nutrient deficiency and other stressors are reflected in the measured spectra (Campolo et al., 2022). For example, the satellite-derived Normalized Difference Vegetation Index (NDVI) can quantify vegetation by measuring the normalized difference between reflectance in near-infrared and red spectral bands (Figure 1-2). The reason is that vegetation can strongly reflects near-infrared lights while mostly absorbing red lights (Feng et al., 2020). Based on the vegetation, the growing states of crops can be assessed, and their yield can be further predicted.

Figure 1-2. A map of NDVI, which can quantify vegetation on the land. The greener, the higher the vegetation is.

As such, several ML models have been explored for crop yield prediction using satellite images at regional scales. For instance, Mkhabela et al. (2011) built linear regression models with the 10-day composite Normalized Difference Vegetation Index (NDVI) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) to predict yield for multiple crops on the Canada Prairies. Bolton and Friedl (2013) derived multiple vegetation indices (VIs) from MODIS data, and linear regression models were then developed based on the extracted VIs for soybean and corn yield prediction across the U.S. corn belt. To assess yield using data from more than one source, non-linear ML models have been established to capture the complex relationships between yield with multimodal features. For example, Johnson (2014) built tree-based regression models to assess sequential remotely sensed VIs and weather variables on soybean and corn yield prediction over the U.S. corn belt and demonstrated the feasibility of using multi-source data for crop production estimation. Kamir et al. (2020) built a support vector regression (SVR) model with time-series MODIS satellite images and weather records for wheat yield prediction, and the model explained 73% of the yield variability across the Australian

wheat belt. Furthermore, Wang et al. (2020) compared the performances between two linear regression models and four non-linear ML models in predicting winter wheat yield in the U.S., and the results showed that the non-linear ML models significantly outperformed the linear ones. Similarly, Chen et al. (2021) integrated satellite imagery, climate data, and meteorological indices for corn yield prediction at the city level in China using four ML approaches, including decision tree-based Cubist, random forest (RF), SVR, and extreme gradient boosting (XGBoost). Chen et al. (2022) proposed a spatial disaggregation method based on several ML methods for corn yield prediction in China at the municipal level.

1.2 Supervised Deep Learning

With the development of artificial neural networks (NN) and the improvement of computing power, deep learning (DL), as a branch of ML, has made impressive progresses in a variety of tasks, including image recognition, autonomous driving, speech recognition, machine translation, and medical diagnosis (Goodfellow et al., 2016). The core idea of DL is to simulate neurons in human brains through fully-connected layers (LeCun et al., 2015). Currently, most DL models are trained through supervised learning, which learns a function to associate the input (i.e., an image or feature vectors) with the label (i.e., response variables) (Russell and Norvig, 2002). Specifically, during training, the input is first fed into the NN, and a prediction is made. After that, a training loss can be calculated based on the prediction and the ground-truth label. The trainable weights in the NN are then updated to minimize the loss during backpropagation. Before the NN is well trained, the training process is normally repeated several times until convergence.

Recently, several supervised DL models have been explored for crop yield prediction (Kang et al., 2020; Ma et al., 2019; You et al., 2017; Yuan et al., 2020). For example, a fully

connected multi-layer perceptron (MLP) was developed by Khaki and Wang (2019) to predict the yield of maize hybrids at field scales in the U.S using soil and weather variables, achieving a root mean square error (RMSE) of 0.86 tons per hectare (t/ha). Besides fully connected MLP, some studies also use more advanced model structures for crop yield prediction. For instance, using time-series MODIS data, You et al. (2017) applied the Gaussian process on two DL architectures, Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), for county-level soybean yield prediction. Their models outperformed USDA prediction by 15% on average. Jiang et al. (2019) built a phenology-based LSTM model for rain-fed corn yield prediction at the county level in nine U.S. Midwestern states. Using the time-series wide dynamic ranged vegetation index (WDRVI) and meteorological variables as input features, an RMSE of 0.87 mg/ha was achieved. Schwalbert et al. (2020) developed an LSTM model for municipality-level soybean yield prediction in Brazil using satellite imagery and climate data and achieved a mean absolute error (MAE) of 0.24 mg/ha. Kang et al. (2020) predicted county-level maize yield in twelve Midwestern states in the U.S. using several ML and DL approaches, including SVR, XGBoost, RF, LSTM, and CNN with an average RMSE of 1.00 t/ha.

Despite the success, existing models mainly focus on predicting the yield without providing any uncertainty information that is important for quantifying the confidence interval of the prediction, which is very useful for practical applications. Also, the plain NN with point estimations on weights typically require an abundance of training data for model training and thus are subject to overfitting. These two issues raise our first research question **(RQ-1): How to quantify predictive uncertainty and increase robustness for supervised deep learning models on corn yield prediction?**

1.3 Unsupervised Domain Adaptation

DL models are data-driven, which means that a large number of labeled training samples (e.g., RS images together with the ground-collecting crop yield records) are required for model training. Therefore, for regions without historical yield records, it is impossible to train a ML model from scratch. Also, most supervised DL models are location-specific or domain-specific. Due to the phenomenon known as domain shift, i.e., data distributions are different in the training region and the testing region, supervised DL models trained within one domain tend to experience a significant decrease in performance when directly applied to a new domain (Kouw and Loog, 2018). For example, RS images in different regions can have different statistical characteristics due to spatial heterogeneity in meteorological conditions, soil properties, and farming practices (Figure 1-3). Therefore, ML models established between reference (reported) yields and RS measurements within a specific spatial domain (i.e., source domain) often lose their validity when directly applied to a new spatial domain (i.e., target domain).

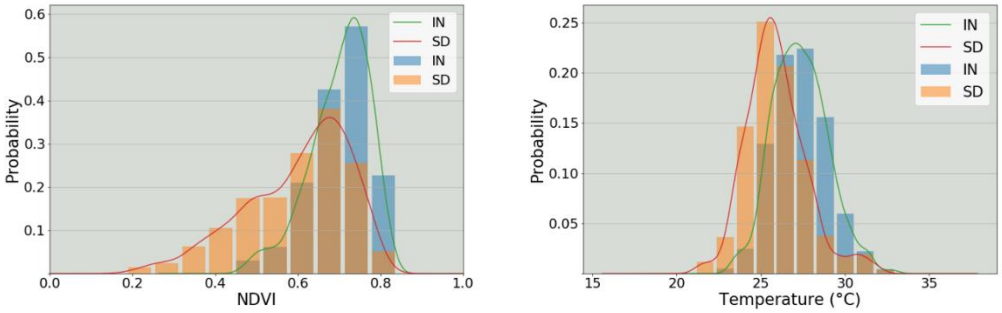

Figure 1-3. Distributions of mean county-level NDVI (left) and mean air temperature (right) during August in Indiana (IN) and South Dakota (SD).

To improve transferability for ML and DL models, transfer learning (TL), a ML technique that transfers knowledge learned from one domain to the other, has become a viable solution (Zhuang et al., 2019). To perform TL for deep NN, a widely used strategy is to first pre-train a NN with labeled samples from the source domain and then fine-tune the weights of the

pre-trained NN using some labeled target samples. For example, Wang et al. (2018) adapted a deep CNN model trained with data from Argentina to predict province-level soybean yields in Brazil by fine-tuning the pre-trained model with labeled data from Brazil. Russello (2018) explored TL between two ecoregions in the U.S. for county-level soybean yield estimation by training a CNN model with labeled data from one ecoregion and fine-tuning it in another region. When an insufficient number of labeled data samples are available in the target domain, there would be concerns about overfitting if fine-tuning the entire network (Mehdipour Ghazi et al., 2017). Therefore, some studies freeze the weights of earlier layers of a pre-trained DL architecture and customize the model to a given task by fine-tuning the last few layers. This framework works based on the idea that earlier layers of a NN learn generic features that can be used in relevant domains (Yosinski et al., 2014). For example, Barbedo (2018) fine-tuned the last few layers of a pre-trained GoogleNet with a plant disease database and adapted the network for plant disease classification. Abdalla et al (2019) fine-tuned a well-trained VGG16 with a small oilseed rape images dataset to classify plants in fields with high-density weeds. Similarly, Chen et al. (2020) freeze the learned weights in the top layers of a well-trained VGG19 and fine-tuned the last three layers with crop disease images to customize the trained model for disease classification.

Despite several successful cases (Barbedo, 2018; Wang et al., 2018), a certain number of labeled data from the target domain is still needed to fine-tune the pre-trained networks. Since collecting crop yield data can be financially expensive, labor-intensive, and time-consuming, many agricultural production areas may lack reliable ground reference yield data for either directly training a ML model or fine-tuning supervised TL models. To improve transferability for ML models without relying on labeled data samples from the target domain, unsupervised

domain adaptation (UDA) has been proposed. The core idea of UDA is to reduce the domain shift between the source domain and the target domain by extracting cross-domain features (Figure 1-4). With the extracted cross-domain features, a cross-domain predictor can be trained and make accurate yield predictions in the target domain.

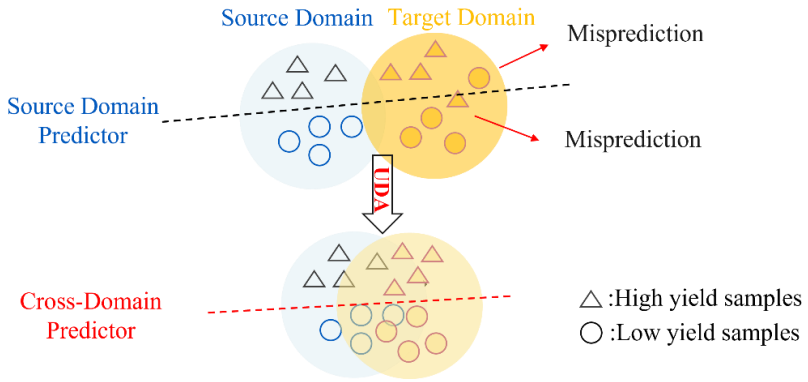

Figure 1-4. A conceptual example of unsupervised domain adaptation (UDA).

Widely used UDA methods can be categorized into two types, including discrepancy-based methods and adversarial-based methods. Discrepancy-based methods try to align features from source and target domains by minimizing the distance between feature distributions (Long et al., 2015; Luo et al., 2017). Adversarial-based methods address the domain shift by learning good representations that are informative for the main learning task and indiscriminative between source and target domains (Ganin et al., 2016).

Existing UDA methods mainly focus on the single-source scenario, i.e., labeled data samples are assumed to be from one source domain. Single-source UDA algorithms commonly employ a conjugated architecture with two objectives (Zhao et al., 2020). One objective is to learn a task model based on the labeled source samples by minimizing the corresponding task losses, such as mean square error loss (MSE) for regression (Feng et al., 2021) and cross-entropy loss for classification (Wang et al., 2021). The other objective is to reduce the domain shift and align the source domain and the target domain. One of the most representative single-source

UDA methods is the domain adversarial neural networks (DANN) (Ganin et al., 2017), which employs an adversarial objective with a domain discriminator to extract domain-invariant features from the source domain and the target domain. However, to our best knowledge, there are no UDA studies conducted for yield prediction which is a regression task that differs from classification applications. This raises our second research question (**RQ-2**): **How does the strategy of adversarial learning can be leveraged to conduct unsupervised domain adaptation on corn yield prediction based on remote sensing and weather observations?**

Although single-source UDA methods have achieved satisfactory results in some real-world applications, the assumption of a single source domain can be invalid in other scenarios when labeled data are from different domains (Zhao et al., 2020). Recently, there has been growing interest in multi-source UDA. Recent multi-source UDA models are mostly developed by extending existing single-source UDA strategies. For example, Peng et al. (2019) proposed a multi-source UDA model named Moment Matching for Multi-Source Domain Adaptation (M^3SDA) for image classification. M^3SDA reduces source-target divergence and inter-source divergence by minimizing the moment-related distance between each domain. Zhao et al. (2018) extended the DANN model and proposed multi-source domain adversarial networks (MDAN) by designing source-specific domain classifiers to realize multi-source UDA. Xu et al. (2018) proposed a deep cocktail network that uses multi-way adversarial learning to minimize the discrepancy between the target and source domains. Zhao et al. (Zhao et al., 2019) designed separate feature extractors for each source and thus could learn more discriminative target representations in an adversarial manner. Tasar et al. (2020) proposed a StandardGAN which standardizes multiple source domains and target domains for satellite image segmentation. Wang et al. (2022) designed domain-specific feature extractors and proposed a multi-source UDA for

unsupervised crop type mapping based on Sentine-2 images. Currently, there are no multi-source UDA studies for agricultural applications such as crop yield prediction. It raises the third research question (**RQ-3**): **How does the strategy of multi-source unsupervised domain adaptation can be leveraged to conduct unsupervised domain adaptation on corn yield prediction based on RS and weather observations?**

1.4 Research Activities

1.4.1 Yield Prediction and Uncertainty Analysis

As mentioned before, existing supervised learning models are unable to provide uncertainty information and are prone to overfitting. These two issues raise **RQ-1: How to quantify predictive uncertainty and increase robustness for supervised deep learning models on corn yield prediction?** Bayesian neural networks (BNN), which introduce Bayesian inference over the weights in the neural networks (Blundell et al., 2015) have provided opportunities to address these issues. Specifically, through Bayesian inference, BNN estimates the predictive distribution rather than a single value. Based on that, the uncertainty interpretation of the target value can be obtained through the spread of the distribution. Furthermore, by adding the prior data distribution in the model development, BNN is less prone to overfitting due to the prior regularization (Gal et al., 2017; LeCun et al., 2015; Nasrabadi, 2007).

To answer RQ-1, this dissertation proposes research activity 1 **(RA-1)**: **Bayesian neural networks for corn yield prediction and uncertainty analysis**. By applying Bayesian inference, the proposed BNN model can provide not only accurate yield estimation but also the corresponding predictive uncertainty. Specifically, informative variables including time-series VIs, sequential weather variables, and soil properties, were first extracted from multiple data sources and aggregated to the county level. A BNN yield prediction model was then developed

based on the extracted features and observed yield records. The proposed model was evaluated in the U.S. corn belt and compared with several state-of-art ML and DL models. Finally, the spatial patterns of the predictive uncertainty were analyzed and the potential driving factors for such patterns were investigated.

1.4.2 Single-source UDA for Corn Yield Prediction

The proposed BNN yield prediction model requires labeled data samples for training and may have low spatial transferability. The success of DANN in applications such as image classification raises the second research question (**RQ-2**): **How does the strategy of adversarial learning can be leveraged to conduct domain adaptation on corn yield prediction based on remote sensing and weather variables?** The DANN consists of three parts, including a feature extractor, a domain classifier, and a label predictor. During the training process, on the one hand, the feature extractor is trained collaboratively with the label predictor to minimize the prediction loss with the aim of extracting task-informative features. On the other hand, the feature extractor is trained adversarially against the domain classifier to maximize the domain loss with the aim of extracting domain-invariant features. As a result, the feature extractor will be updated towards extracting task-informative and domain-invariant features to help alleviate the negative impact of domain shift and make accurate predictions across two domains.

Despite success in classification applications, DANN cannot be directly applied for regression tasks such as crop yield prediction. This is mainly because the predefined weighting parameter in the original DANN model, which controls the trade-off between the prediction loss and the domain loss, needs to be adjusted according to the yield magnitudes which can change dramatically in different agricultural production regions as well as in different harvest years. Also, since the training set for county-level corn yield prediction is comparatively small,

overfitting may happen during the training of DANN. To address these issues, the dissertation proposes research activity 2 (**RA-2**): **Adversarial domain adaptation on corn yield prediction,** in which two variants of DANN, i.e., Adaptive DANN (ADANN) and Bayesian DANN (BDANN), are developed for corn yield prediction at the county level based on remote sensing images and weather variables. The ADANN model was designed to adaptively adjust the weighting parameter between the yield prediction loss and the domain classification loss. Based on ADANN, we further applied Bayesian inference to model training and designed the BDANN model. Both models were evaluated in two ecoregions of the U.S. corn belt and compared with other widely used ML and DL models.

1.4.3 Multi-source UDA for Corn Yield Prediction

Although single-source UDA is a promising solution to improve transferability for DL models with no need for reference yield records in the target domain, there are still two major bottlenecks in applying UDA methods to crop yield prediction based on remote sensing images and weather variables. First, most current UDA methods are designed for single-source UDA with the assumption that all labeled samples are from the same domain. In practice, labeled training samples may be collected from multiple source domains with different feature distributions (Zhao et al., 2020). Since domain shifts exist not only between source and target but also among different source domains, single-source UDA methods could have a poor performance when samples from different sources interfere with each other (Riemer et al., 2019). Second, current UDA methods mostly align distributions of source and target without considering specific tasks. For remote sensing images and weather variables with significant domain shifts, directly aligning feature distributions may project the data into ambiguous feature

spaces with no meaningful information. As a result, misalignment would happen and misprediction would still be made.

Recently, there have been growing interests in multi-source UDA to address the aforementioned issues (Lin et al., 2020; Zuo et al., 2021). However, there are no multi-source UDA studies for agricultural applications such as crop yield prediction. This raises our third research question (RQ-3): **How does the strategy of multiple unsupervised domain adaptation can be leveraged to conduct domain adaptation on corn yield prediction based on remote sensing and weather observations?**

To answer RQ-3, this dissertation proposed research activity 3 (**RA-3**): **Multi-source Maximum Predictor Discrepancy (MMPD) for unsupervised domain adaptation on corn yield prediction.** We used the strategy of multi-source UDA to group labeled data samples into multiple sources and aligned the target domain to each source domain separately. Also, the idea of maximum predictor discrepancy (MPD) was used to conduct UDA by considering specific tasks via pairs of domain-specific yield predictors. As such, data samples from different regions were grouped into multiple sources for multi-source domain adaptation. Then, by using MPD, the MMPD model tried to align the distributions of source and target domains by considering task-specific regression models. The proposed MMPD model was evaluated in three scenarios in the U.S. corn belt and Argentina and compared with other state-of-art UDA methods.

1.5 Organization of the Dissertation

There are five chapters in this dissertation. Chapter 1 introduces the background of this dissertation and raises the research questions followed by the investigated RAs. Chapter 2 presents the study areas and the data sources for this study. Chapter 3 proposes the BNN model

for county-level corn yield prediction and uncertainty analysis. Chapter 4 develops ADANN and BDANN for UDA on corn yield prediction at the county level. Chapter 5 proposes the MMPD model for multi-source UDA on corn yield prediction at the county level. Chapter 6 concludes the major contributions and limitations of the dissertation and discusses potential future work. A list of peer-reviewed publications and manuscripts in preparation corresponding to major chapters of the dissertation are summarized as follows.

Chapter 3:

- Ma, Y., Zhang, Z., Kang, Y. and Özdoğan, M., 2021a. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259, p.112408.

Chapter 4:

- Ma, Y., Zhang, Z., Yang, H.L. and Yang, Z., 2021b. An adaptive adversarial domain adaptation approach for corn yield prediction. *Computers and Electronics in Agriculture, 187, p.106314*.
- Ma, Y. and Zhang, Z., 2022. A Bayesian Domain Adversarial Neural Network for Corn Yield Prediction. *IEEE Geoscience and Remote Sensing Letters, 19, pp 1-5*.

Chapter 5:

- Ma, Y., Zhang, Z, 2022. Multi-source Unsupervised Domain Adaptation on Corn Yield Prediction. *AAAI-22 AI for Agriculture and Food Systems (AIAFS) Workshop*.
- Ma, Y., Zhang, Z., 2022. Maximum Predictor Discrepancy for Multi-source Unsupervised Domain Adaptation on Corn Yield Prediction (*Manuscript Submitted*).

CHAPTER 2 MATERIALS

The U.S. corn belt and Argentina are both the top corn-producing regions in the world. They were selected as the study areas due to the availability of sufficient yield records for model development and validation. Feature variables were collected from multiple data sources and paired with the corresponding county-level yield records for model development.

2.1 Study Areas

The study areas include the U.S. corn belt and Argentina. The U.S. corn belt is in the Midwestern United States (Figure 2-1). This area is geographically flat and relatively flat with fertile soils (Johnson, 2014). Therefore, It has become the main agricultural region in the U.S. since the year 1850 and accounts for over 75% of the annual corn production in the U.S. (USDA, 2020b). County-level historical yield records in the U.S. corn belt from 2001 to 2019 were collected from the USDA National Agricultural Statistics Service (NASS) Quick Stats Database, a platform for accessing U.S. agricultural data (USDA, 2020b).

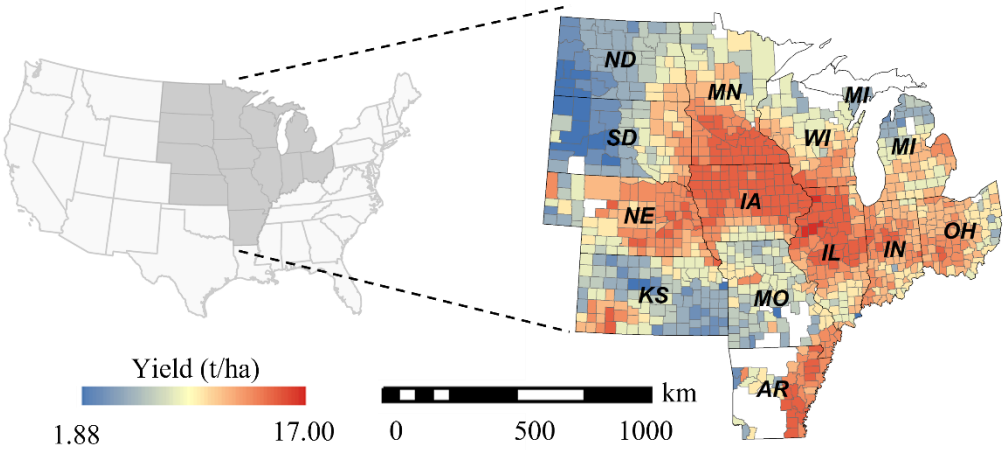

Figure 2-1. The county-level average yield of corn in the U.S. corn belt over the years 2008–2019.

Argentina is also a significant producer and exporter of corn. Being in the Southern Hemisphere, seasons in Argentina are the reverse of those of the U.S. Cornfields in Argentina

mostly locate in several provinces (Figure 2-2), including Buenos Aire (BA), Santa Fe (SF), and Santiago del Estero (SE). This region has plenty of rainfall and thus has a favorable climate for rainfed crop production (Global Yield Gap Atlas, 2021). Historical yield records at the county level in Argentina corn production areas from 2006 to 2019 were collected from the online platform by the Argentina Ministry of Agriculture (Argentine Undersecretary of Agriculture, 2020), which provides historical planted acreage, harvested acreage, production, and yield for main crops.

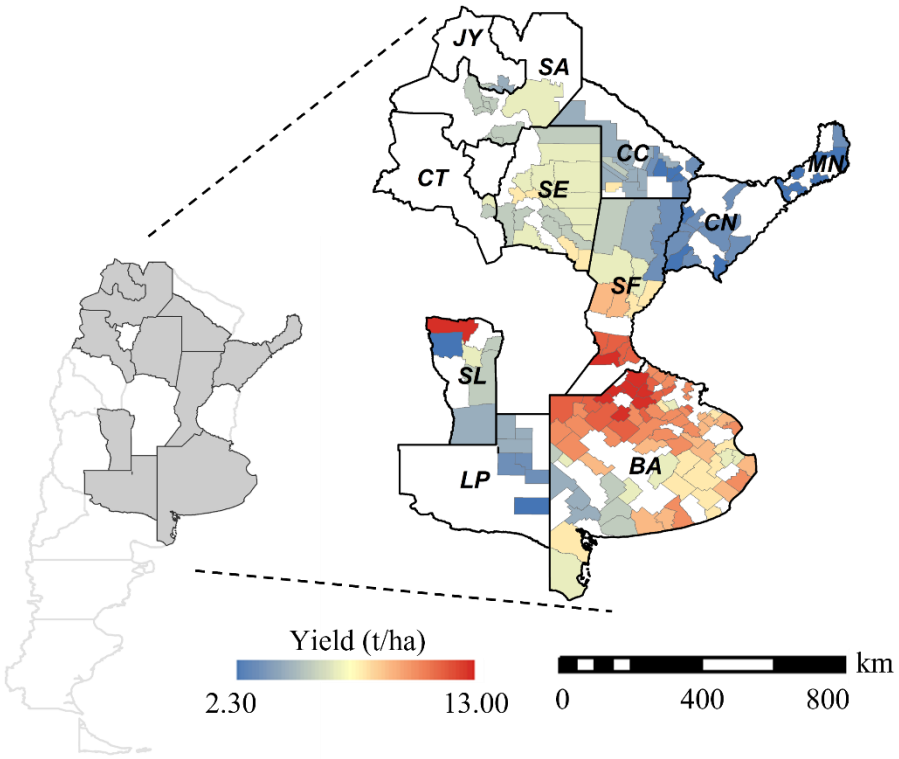

Figure 2-2. The county-level average yield of corn in Argentina over the years 2008–2019.

2.2 Satellite Imagery

Since the VIs have been widely used for yield prediction (Bolton and Friedl, 2013; Johnson, 2014), they were considered in this study and extracted from the daily MODIS MCD43A4 product (Table 2.1) which provides visible, near-infrared (NIR), and shortwave

infrared (SWIR) reflectance data at 500-m spatial resolution (Schaaf and Wang, 2015). The MODIS MCD43A4 product is generated daily using 16 days of Terra and Aqua MODIS data (Wang et al., 2020).

Table 2-1. Spectral band information of the MODIS MCD43A4 product.

Band	Wavelength (nm)	Name
1	620-670	Red
2	841-876	NIR
3	459-479	Blue
4	545–565	Green
5	1230–1250	SWIR
6	1628–1652	SWIR2
7	2105–2155	SWIR3

Based on the spectral bands from MODIS MCD43A4 product, three complementary VIs were calculated and used as predictors, including Enhanced Vegetation Index (EVI), Green Chlorophyll Index (GCI), and Normalized Difference Water Index (NDWI). Specifically, EVI is an enhanced version of NDVI that has higher sensitivity in high biomass regions and can more precisely quantify vegetation on the ground (Gao et al., 2000; Huete et al., 2002). GCI quantifies the light use efficiency by measuring the canopy chlorophyll content and can be used as an indicator for crop health (Gitelson et al., 2005). NDWI quantifies the vegetation moisture content, and thus it is widely utilized to measure the water content changes in crop leaves and can monitor droughts (Bolton & Friedl, 2013; Gao, 1996). These three VIs were calculated as below (Eq. (2.1-2.3)):

$$EVI = 2.5 \times \frac{(NIR - Red)}{NIR + C_1 \times Red - C_2 \times Blue + L} \quad (2.1)$$

$$GCI = \frac{NIR}{Green} - 1 \quad (2.2)$$

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR} \quad (2.3)$$

where Red, $Blue$, $Green\ NIR$, and $SWIR$ respectively represent the atmospherically corrected surface reflectance in red, blue, green, near-infrared, and shortwave infrared channels; L stands for the soil and canopy background adjustment factor; C_1 and C_2 denote the coefficients to correct atmospheric influences. The coefficients in Eq. (2.2) were set to L = 1, C_1 = 6, and C_2 = 7.5 according to the MODIS EVI algorithm from previous studies (Jiang et al., 2008; Kang et al., 2020).

Besides satellite VIs, daytime and nighttime land surface temperature (LSTday and LSTnight) were collected and extracted from the MODIS MYD11A2 product, which provides daily data with a 1-km spatial resolution (Park et al., 2005). These two variables were considered because they have been used for monitoring agricultural drought, which is a critical environmental stressor that can significantly affect crop productivity (Guan et al., 2017; Johnson, 2014; Lobell et al., 2013).

2.3 Weather Variables

The weather variables were extracted from the dataset generated by the Parameter elevation Regressions on Independent Slopes Model (PRISM), which is a climate analysis system that uses point data, a digital elevation model (DEM), and other spatial datasets to generate gridded estimates of climatic parameters (Daly et al., 2008). To be concrete, six primary meteorological variables, including daily mean, minimum, maximum temperature (Tmean, Tmin, and Tmax), daily minimum, and maximum Vapor Pressure Deficit (VPDmin, VPDmax), and daily total precipitation (PPT), were collected from the PRISM dataset which has a 4-km spatial resolution. Besides, we also used the Global Land Data Assimilation System (GLDAS)

dataset which generates optimal fields of land surface states and fluxes at 0.25 arc degree spatial resolution by ingesting satellite- and ground-based observational data products and using advanced land surface modeling and data assimilation techniques (Rodell et al., 2004). Two variables were derived from the GLDAS dataset. One was the daily mean evapotranspiration (ET) which is related to the atmospheric water cycle, and the other was a water stress indicator (GLDASws) calculated as the ratio of ET and potential ET collected from the GLDAS dataset (Kang et al., 2020).

Considering that PRISM only covers the conterminous United States and is not available for Argentina, we extracted PPT from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset with a resolution of ~5.5 km (Funk et al., 2015) and Tmean, Tmax, and Tmean from the ERA5 reanalysis dataset with a resolution of ~0.25 arc degrees for counties in Argentina (Cunha and Silva, 2020). CHIRPS is a long-term quasi-global rainfall dataset, which incorporates satellite imagery with ground station data to generate gridded time-series rainfall for global drought monitoring (Funk et al., 2015). ERA5 combines climate model data with observations from across the world for atmospheric reanalysis of the global climate (Cunha and Silva, 2020).

2.4 Soil Properties

Soil properties are also critical for plant growth and have significant impacts on crop yield. Three types of soil properties in the U.S. corn belt were collected, including Soil Available Water Holding Capacity (AWC), Soil Organic Matter (SOM), and Cation Exchange Capacity (CEC), which were derived at 30-m spatial resolution from Soil Survey Geographic database (SSURGO) (Soil Survey Staff et al., 2020). AWC quantifies the water availability in soil, which directly influences root and plant growth. SOM represents the amount of soil organic matter, and

higher SOM can help reduce soil erosion rates and increase water and nutrient retention. CEC is included since it measures the capacity of a soil to hold essential nutrients and thus a good indicator of the soil's potential to harbor a healthy crop.

2.5 Data Preprocessing

Google Earth Engine (GEE) platform was leveraged to preprocess the data. Data collection in the U.S. corn belt started from 2001 to 2019 while data collection in Argentina started from 2006 to 2019. Specifically, spatial filtering was first conducted by using the MODIS Land Cover Type product (MCD12Q1 v6) at 500-m spatial resolution and NASS Cropland Data Layer (CDL) as the crop masks to mask out observations on non-cultivated croplands in each county. After that, each type of variable was aggregated spatially at the county level by calculating the mean value. Then, sequential variables including VIs and weather variables were aggregated into a 16-day interval to cover the complete planting and growing season for corn. Finally, the feature variables were paired with county-level yield records and used for model development.

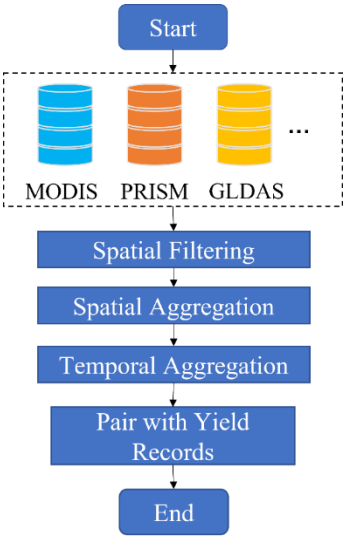

Figure 2-3. The data preprocessing steps.

CHAPTER 3 BAYESIAN NEURAL NETWORKS FOR CORN YIELD PREDICTION AND UNCERTAINTY ANALYSIS

3.1 Overview

Accurate and timely prediction of corn yield is essential for regional food security. Traditional ML and DL methods are trained to associate input feature vectors with yield records but cannot quantify the predictive uncertainty. To achieve accurate corn yield prediction and simultaneously quantify its uncertainty, a BNN was proposed. By applying Bayesian inference, the BNN yield prediction model can estimate the predictive distribution and thus provide both yield prediction and corresponding predictive uncertainty through the spread of the distribution. Furthermore, by adding the prior data distribution in the model development, BNN is less prone to overfitting due to the prior regularization (Gal et al., 2017; LeCun et al., 2015; Nasrabadi, 2007).

Feature variables were collected from multiple data sources, including remote sensing (RS) imagery, weather variables, soil properties, and historical average yield. Experiments in the U.S. corn belt in 2010-2019 showed that the BNN model outperformed state-of-art ML and DL models. Also, the in-season prediction performance of the BNN model was evaluated within the growing season from May to October at a 16-day interval. The optimum prediction accuracy was achieved by the BNN model by the middle of August, which is about two months before the harvest. Besides, we also analyzed the predictive uncertainty and found that more than 84% of the observed yield records were successfully enveloped in the 95% confidence interval of the

predictive yield distribution. Finally, the potential driving factors for the predictive uncertainty were investigated and several factors were found to be highly correlated with the predictive uncertainty, including the observation noise and the environmental stress such as heat and water stress.

3.2 Methodology

3.2.1 Fundamentals of BNN

A NN can be viewed as a probabilistic model $p(y|\mathbf{x}, \mathbf{w})$ to associate the input $\mathbf{x}$ with the response variable y through successive hidden layers with weights $\mathbf{w}$. For yield prediction, which is a regression problem, y is a continuous variable, and $p(y|\mathbf{x}, \mathbf{w})$ is assumed to be a Gaussian distribution (Nasrabadi, 2007). Given a data sample $(\mathbf{x}_i, y_i)$, the NN is trained to predict the yield distribution $N(\hat{y}_i, \hat{\sigma}_i^2)$:

$$p(y_i|x_i, \widehat{\mathbf{w}}^{(t)}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\hat{\sigma}_i^2}\right) \tag{3.1}$$

Correspondingly, given training samples $\mathcal{D} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \dots, (\mathbf{x_N}, y_N)\}$, the NN is trained through the maximum likelihood estimation (MLE) (LeCun et al., 2015). In practice, the NN is commonly updated to minimize the negative log-likelihood function (Eq. (3.2)). Moreover, in a traditional NN for regression, it is assumed that the $p(y|\mathbf{x}, \mathbf{w})$ has a constant standard deviation. As a result, the loss function can be converted to minimize the mean squared error (MSE) (Eq. (3.3)):

$$\begin{aligned} -\log p\left(\mathcal{D}\middle|\widehat{\mathbf{w}}^{(t)}\right) &= -\log \textstyle\prod_{i=1}^{N} p(y_i|x_i, \widehat{\mathbf{w}}^{(t)}) \\ &= -\log \textstyle\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\hat{\sigma}_i^2}\right) \\ &= -\textstyle\sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\hat{\sigma}_i^2}\right) \end{aligned} \tag{3.2}$$

$$= \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{2\hat{\sigma}_i^2} + \sqrt{2\pi\hat{\sigma}_i^2}$$

$$MSE = \frac{1}{N}(y_i - \hat{y}_i)^2 \tag{3.3}$$

Though successful, the traditional NN was trained to do point estimation on each trainable weight. It means that each trainable weight is estimated to be a single value. As a result, it makes NN subject to overfitting, especially when the training size is relatively small compared to the number of trainable parameters (Blundell et al., 2015; Deodato et al., 2019). More importantly, since each weight is fixed after training, a well-trained NN can only predict the yield without quantifying the predictive uncertainty, which is very important for model evaluation.

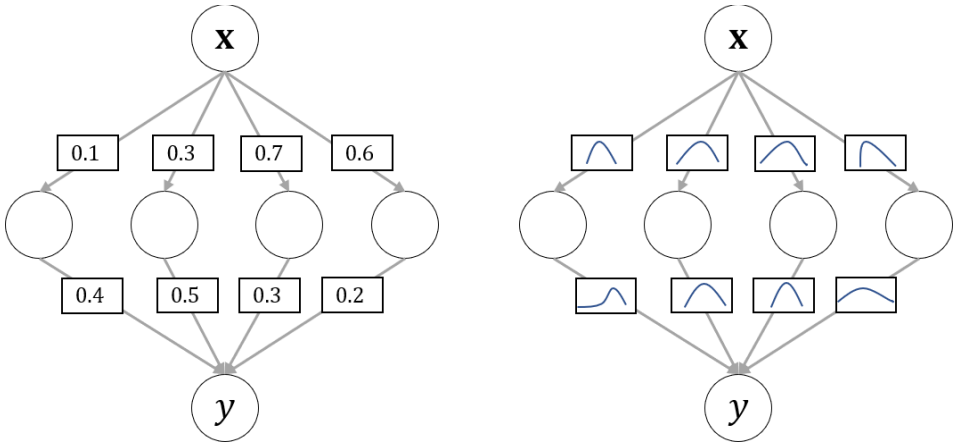

Figure 3-1. Comparison between traditional and Bayesian NN. Left: traditional NN with each weight modeled as a fixed value. Right: BNN with each weight modeled as a probability distribution.

To improve the DL models' robustness against overfitting and quantify predictive uncertainty, BNN was proposed by applying Bayesian inference to the traditional NN. Instead of having point estimation on weights, all weights in BNN are represented by probability distributions, and the difference between traditional NN and BNN is illustrated in Figure 3-1. Specifically, the posterior distribution $p(\mathbf{w}|\mathcal{D})$ of the weights is estimated according to the Bayes theorem. For an input sample **x**, its predictive distribution of label y can then be made by taking

an expectation under $p(\mathbf{w}|\mathcal{D})$, which is denoted as $p(y|\mathbf{x},\mathcal{D}) = \int p(y|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}$. However, it is intractable to directly estimate posterior distributions for weights considering the size of deep NN. Instead, the posterior on the weights are approximated by a variational distribution via variational learning (Blundell et al., 2015), which finds the variational parameters $\boldsymbol{\theta}$ of a distribution on the weights $q(\mathbf{w}|\boldsymbol{\theta})$ that minimize the KL-divergence with the posterior on the weights $p(\mathbf{w}|\mathcal{D})$:

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w}|\mathcal{D})] \\ &= \arg\min_{\boldsymbol{\theta}} KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w})] + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}[-\log p(\boldsymbol{\mathcal{D}}|\mathbf{w})]\end{aligned} \tag{3.4}$$

There are two terms in Eq. (3.4): The first term is a prior-dependent loss $L_p = KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w})]$, which is the KL-divergence between the variational distribution $q(\mathbf{w}\,|\,\boldsymbol{\theta})$ and the prior distribution $p(\mathbf{w})$ of trainable parameters $\mathbf{w}$. The second term, $E_{q(\mathbf{w}|\boldsymbol{\theta})}[-\log p(\boldsymbol{\mathcal{D}}|\mathbf{w})]$, is a data-dependent loss in the form of a negative log-likelihood function. Use Monte Carlo integration and sample $\widehat{\mathbf{w}}^{(t)}$ from $q(\mathbf{w}|\boldsymbol{\theta})$, the loss function can be approximated as:

$$\begin{aligned}\mathcal{F}(\mathcal{D},\boldsymbol{\theta}) &\approx \frac{1}{T}\sum_{t=1}^{T}\left[\log q\left(\widehat{\mathbf{w}}^{(t)}\middle|\boldsymbol{\theta}\right) - \log p\left(\widehat{\mathbf{w}}^{(t)}\right) - \log P\left(\mathcal{D}\middle|\widehat{\mathbf{w}}^{(t)}\right)\right] \\ &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\left[\log q\left(\widehat{\mathbf{w}}^{(t)}\middle|\boldsymbol{\theta}\right) - \log p\left(\widehat{\mathbf{w}}^{(t)}\right) - \sum_{i=1}^{N}\log P\left(y_i\middle|x_i,\widehat{\mathbf{w}}^{(t)}\right)\right]\end{aligned} \tag{3.5}$$

in which (x_i, y_i) is the i-th data samples; N is the total number of training samples; T is the total number of sampling times; In practice, we can just sample once (e.g., T=1).

After training, every possible configuration of $\mathbf{w}$ sampled from $q(\mathbf{w}|\boldsymbol{\theta})$ can make a prediction, and taking the expectation is equivalent to average predictions using an ensemble of NN weighted by the posterior distribution of weights, which can be considered as a special case of ensemble learning (Jospin et al., 2020). Also, regularization is introduced by placing the prior

distribution $p(\mathbf{w})$. These together help prevent overfitting and thus improve the model generalization ability. More importantly, since each weight is modeled as a distribution, the prediction from BNN is different in each run. Therefore, given an input data sample, its predictive uncertainty can be estimated by drawing weight samples from the posterior distribution and running a series of predictions.

3.2.2 Model Architecture

In this study, we aimed to not only predict the corn yield but also obtain the uncertainty associated with the prediction. In a BNN with a single output of yield prediction (Figure 3-1 Right), given an input data sample, its prediction uncertainty can be obtained by drawing weight samples from the variational distributions and running a series of predictions. The uncertainty is estimated by calculating the standard deviation of the series of predictions. Although feasible, this approach cannot directly estimate the predictive uncertainty but requires Monte Carlo sampling which is computationally intensive and time-consuming (Goodfellow et al., 2016).

To enable the model to predict the yield and estimate the predictive uncertainty simultaneously, we designed a BNN architecture with two endpoints (Figure 3-2). As described in Section 3.1, for a probabilistic model, given an input data sample, the predicted yield distribution is in the form of a Gaussian distribution. Therefore, the proposed BNN model is trained to predict the yield distribution with one endpoint outputting the mean of the yield distribution and the other endpoint outputting the standard deviation. Specifically, the developed BNN model starts with a multi-layer feature extraction net to extract high-level features from the inputs. Then, the extracted features are fed into two independent sub-networks, namely “yield net” and “uncertainty net”, to respectively estimate the mean $\hat{y}$ and standard deviation $\hat{\sigma}$ for the

final predictive distribution $N(\hat{y}, \hat{\sigma})$, in which $\hat{y}$ is the unbiased estimation of the predicted yield value and $\hat{\sigma}$ quantifies the predictive uncertainty (i.e., a larger $\hat{\sigma}$ indicates higher uncertainty).

Through experimental analysis, the architecture of the BNN model was designed to have a depth of five (Figure 3-2). The feature extraction net started with an input layer followed by two hidden layers with 256 and 128 neurons respectively, and both the yield net and uncertainty net included two fully connected hidden layers with 64 and 32 neurons for each. We chose the Rectified Linear Unit (ReLU) as the activation function and Adam algorithm was adopted as the optimizer to update trainable parameters after each epoch (Goodfellow et al., 2016). The number of training iterations was set to 1500 epochs with the batch size as 512.

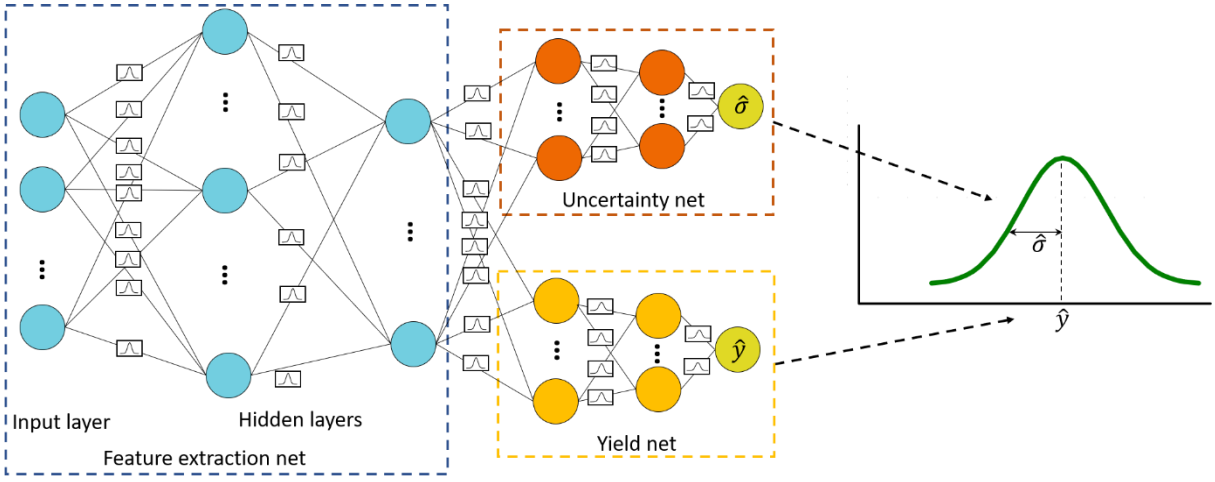

Figure 3-2. The architecture of the developed BNN model.

3.3 Experimental Setup

The BNN model was trained and evaluated within twelve Midwestern U.S. states, including North Dakota, South Dakota, Kansas, Nebraska, Minnesota, Iowa, Wisconsin, Illinois, Indiana, Ohio, Missouri, and Arkansas (Johnson, 2014). Informative variables including time-series VIs, sequential weather variables, soil properties, harvest year, and historical average yield (Table 3.1), were first extracted from multiple data sources and aggregated to the county level. MODIS Land Cover Type product (MCD12Q1 v6) at 500-m spatial resolution was applied as the

mask layer to exclude non-cultivated areas. The sequential feature variables were then aggregated into a 16-day interval from March 10th to October 4th. Finally, the extracted feature variables and yield records were paired for model development. Data collection was from 2001 to 2019.

Table 3-1. Summary of the input features for model development.

Category	Variables	Mask Layer	Related properties	Spatial Resolution	Source	Latency
Satellite Imagery	EVI	MODIS Land Cover Type product (MCD12Q1 v6)	Plant vigor	500 m	MODIS	One day
	GCI					
	NDWI					
	LSTday (Kelvin)		Heat stress	1 km		
	LSTnight (Kelvin)					
Weather	Tmean (°C)			4 km	PRISM	One day
	Tmax (°C)					
	Tmin (°C)					
	PPT (mm)		Water stress			
	VPDmax (hPa)					
	VPDmean (hPa)					
	VPDmin (hPa)					
	GLDASws[1]			0.25 arc degree	GLDAS	One month
	ET (mm)					
Soil	AWC (cm)		Soil water uptake	30 m	SSURGO	N/A
	SOM (kg/m^2)		Soil nutrient uptake			
	CEC (cmol/kg)					
Others	Harvest Year			County-level	USDA NASS	N/A
	Historical average yield (t/ha)					

To validate the developed BNN model, it was compared to five widely used ML models, including (i) three traditional ML models: Ridge regression (Ridge), RF, and SVR; and (ii) two representative DL models: multilayer perceptron (MLP) and LSTM. All the approaches were

evaluated on ten testing years 2010–2019, and for each testing year, data in all preceding years since 2001 were used for model training. To evaluate the performance of each model, three metrics including coefficient of determination (R^2), RMSE, and mean absolute relative error (MARE) were selected as the metrics and calculated as (Eq. (3.5-3.7)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.6}$$

$$\text{MARE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.7}$$

where n denotes the number of data samples; y_i and $\hat{y}_i$ are the reference yield record and estimated yield for the ith county in the testing set; $\bar{y}$ denotes the mean values of reference yield records in the testing set.

3.4 Results and Discussion

3.4.1 End-of-season Prediction Performance

We first compared the end-of-season prediction performance on October 4th with the full-length feature set, and the accuracies of the six models on each testing year over five runs are reported in Table 3.2-3.4. It was observed that the developed BNN model outperformed all the other ML and DL methods with the smallest year-to-year variation, reaching an average R^2 of 0.77 over ten testing years. It was observed that BNN was the best-performing model in most testing years. LSTM is the second best, demonstrating its strong capability of handling time-series features. Also, it was notable that the non-linear ML approaches had much better

performances than the linear Ridge regression model, showing that the linear model could fail to capture the complex relationship between multi-source data and crop yield.

Table 3-2. The R^2 of end-of-season evaluation in 2010-2019.

Year	Ridge	RF	SVR	MLP	LSTM	BNN
2010	0.62	0.66	0.65	0.61	0.64	**0.72**
2011	0.67	0.71	0.68	0.69	0.63	**0.73**
2012	0.47	0.54	0.59	0.60	0.64	**0.70**
2013	0.55	0.67	0.56	0.63	**0.74**	0.72
2014	0.54	0.66	0.61	0.69	0.71	**0.82**
2015	0.43	0.67	0.56	0.72	0.72	**0.75**
2016	0.54	0.65	0.68	0.70	0.72	**0.82**
2017	0.64	0.74	0.75	0.75	0.79	**0.83**
2018	0.65	0.76	0.67	0.77	0.78	**0.82**
2019	0.24	0.48	0.50	0.63	0.65	**0.76**
Average	0.54	0.65	0.63	0.68	0.70	**0.77**

Table 3-3. The RMSE (t/ha) of end-of-season evaluation in 2010-2019.

Year	Ridge	RF	SVR	MLP	LSTM	BNN
2010	1.19	1.13	1.15	1.22	1.16	**1.05**
2011	1.31	1.22	1.28	1.27	1.35	**1.14**
2012	2.07	1.89	1.73	1.69	1.57	**1.41**
2013	1.44	1.24	1.42	1.36	**1.06**	1.13
2014	1.37	1.18	1.26	1.13	1.09	**0.86**
2015	1.52	1.14	1.33	1.06	1.08	**1.01**
2016	1.36	1.18	1.12	1.09	1.04	**0.85**
2017	1.35	1.16	1.14	1.15	1.03	**0.94**
2018	1.41	1.17	1.38	1.14	1.12	**0.98**
2019	1.63	1.34	1.28	1.14	1.07	**0.92**
Average	1.47	1.27	1.31	1.23	1.16	**1.03**

Table 3-4. The MARE (%) of end-of-season evaluation in 2010-2019.

Year	Ridge	RF	SVR	MLP	LSTM	BNN
2010	11.18	10.48	10.54	11.31	10.58	**9.14**
2011	15.45	15.00	15.32	15.18	16.36	**13.65**
2012	34.29	30.70	28.23	28.09	25.64	**22.23**
2013	12.40	10.58	12.20	11.26	**9.26**	9.55
2014	9.84	8.93	9.32	8.01	7.82	**6.46**
2015	11.77	8.78	9.76	8.02	8.01	**7.63**
2016	10.03	8.87	8.39	8.07	7.94	**6.35**
2017	9.72	8.81	8.65	8.68	8.22	**6.83**
2018	10.98	8.48	10.75	8.40	8.32	**7.42**

2019	13.84	9.89	9.72	8.32	8.26	**6.30**
Average	13.95	12.05	12.29	11.53	11.04	**9.56**

To evaluate whether the methods are statistically different on the reported R^2, we used a paired sample t-test to perform the statistical tests between the BNN evaluation results and each comparison model. A t-test is a statistical test that compares the means of two samples (Mishra et al., 2019). In our case, we compared the means of R^2 of each model in all testing years. Since the experiment was repeated five times in each testing year, there were totally 50 pairs of samples in the t-test. As shown in Table 3-5, the accuracy improvement obtained by the proposed BNN model was statistically significant.

Table 3-5 Results of the paired sample t-test between the R^2 of each comparison model and BNN in all testing years.

Model	**t**	***p*-value**
BNN vs Ridge	55.390	0.000
BNN vs. RF	38.892	0.000
BNN vs. SVR	52.580	0.000
BNN vs. MLP	37.676	0.000
BNN vs. LSTM	27.811	0.000

We then presented the density scatter plots of all the methods in Fig 3.5 to further show the agreement between the reported and the predicted yield in three representative testing years, including 2012, 2016, and 2019. 2016 is an average year with a normal climate and normal planting progress. Therefore, 2016 is regarded as a normal year. On the other hand, prolonged heat occurred in the U.S. corn belt during the middle of summer in 2012 and an unusually wet spring followed by an unusually cool June postponed the corn planting in the U.S. corn belt in 2019 (Johnson, 2014; USDA, 2020b), which made these two years abnormal. The best agreement was again observed in the developed BNN model over all three selected testing years (Fig 3.5(f1-f3)), while the linear Ridge regression model showed less agreement than the other

approaches (Fig 3.5(a1-a3)). The proposed BNN model outperformed all the other ML and DL methods and could not only achieve high prediction accuracy with an R^2 over 0.80 in average years (Fig 3.5(f2)) but also had stable performance in abnormal years (Fig 3.5(f1)&(f3)). Besides, it was notable that RF, SVR, and MLP made severe overestimation errors in 2012 (Figs 3.5(b1), (c1), (d1)) and had severe underestimation errors in 2016 and 2019 (Figs 3.5(b2-b3), (c2-c3), (d2-d3)). In 2012, most counties experienced large yield losses due to the prolonged drought. As a result, the trained models tended to overestimate the corn yield in 2012. In 2016 and 2019, due to the imbalanced training set in which there were relatively fewer high-yield samples, models would be trained biasedly and made overestimation errors. Although the LSTM showed improvement in these aspects, large errors were still observed in low-yield counties (Figure 3-5(e1-e3)) in comparison with the proposed BNN model.

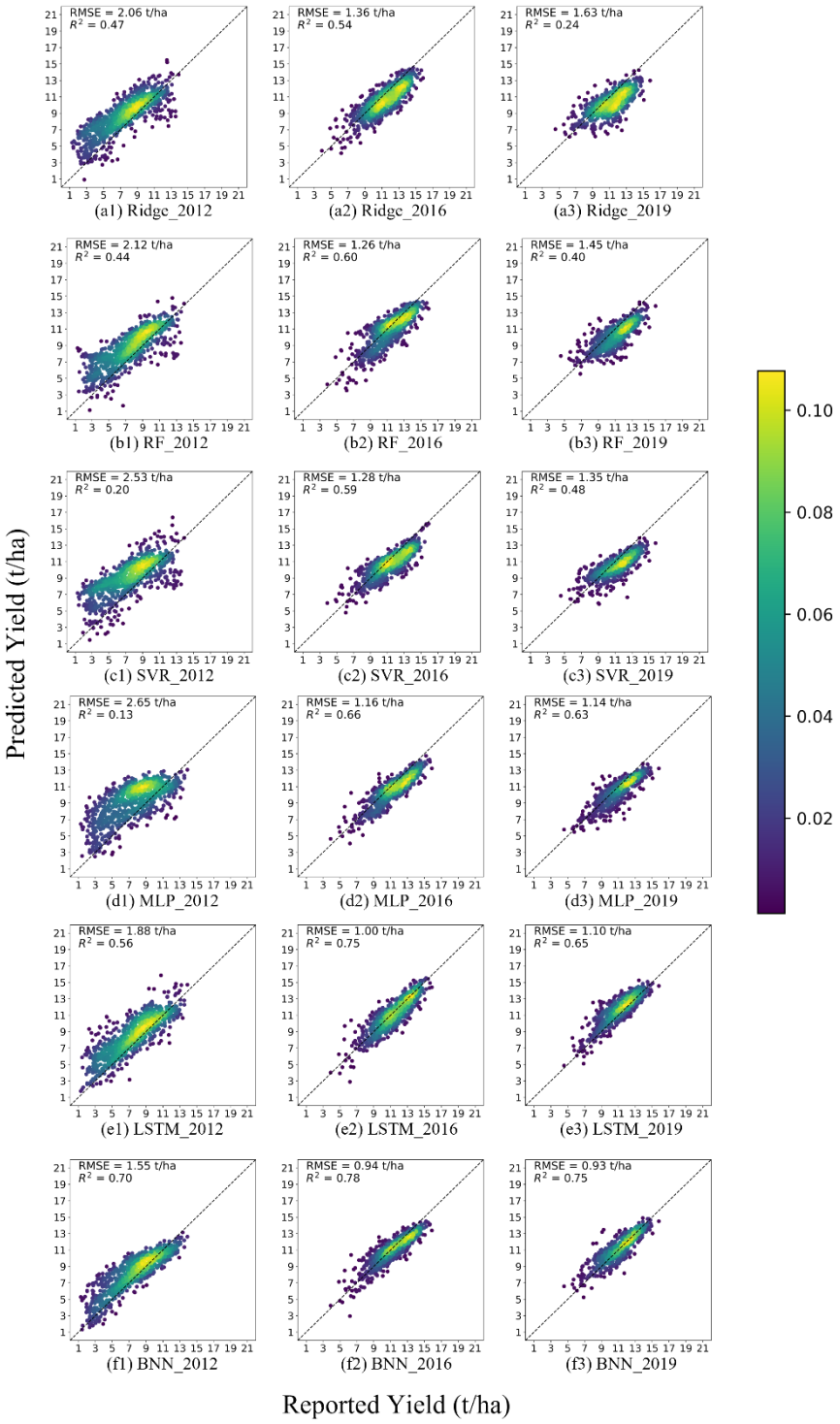

Figure 3-3. The density scatter plots of reported yields vs. predicted yields of (a) Ridge, (b) RF, (c) SVR, (d) MLP, (e) LSTM, and (f) BNN in three testing years: (1) 2012; (2) 2016; (3) 2019.

To further illustrate the performances of different models, we presented the absolute relative error maps in 2012, 2016, and 2019 for each model in Figure 3-4, in which darker color represents a larger error. The results showed that compared to other states, larger errors were observed in North Dakota, South Dakota, Kansas, Arkansas, and Missouri, in all the approaches. This is likely because of environmental stresses (i.e., water stress and heat stress) in these areas, which can introduce uncertainty in crop productivity and increase the difficulties in modeling crop yield variability. Subsequently, comparing different approaches, the BNN model outperformed the other models and had smaller errors for most counties in both the average year-2016 (Figure 3-4 (f2)) and abnormal years-2012 and 2019 (Figure 3-4 (f1)&(f3)). In particular, the linear Ridge regression model performed the worst among the six models and made more errors in Iowa, Wisconsin, and Illinois in 2016 and 2019 (Figure 3-4 (a2)-(a3)) when compared to other methods. Some improvements were shown in all the non-linear models, and errors in the central region were significantly reduced in the prediction results by LSTM and BNN.

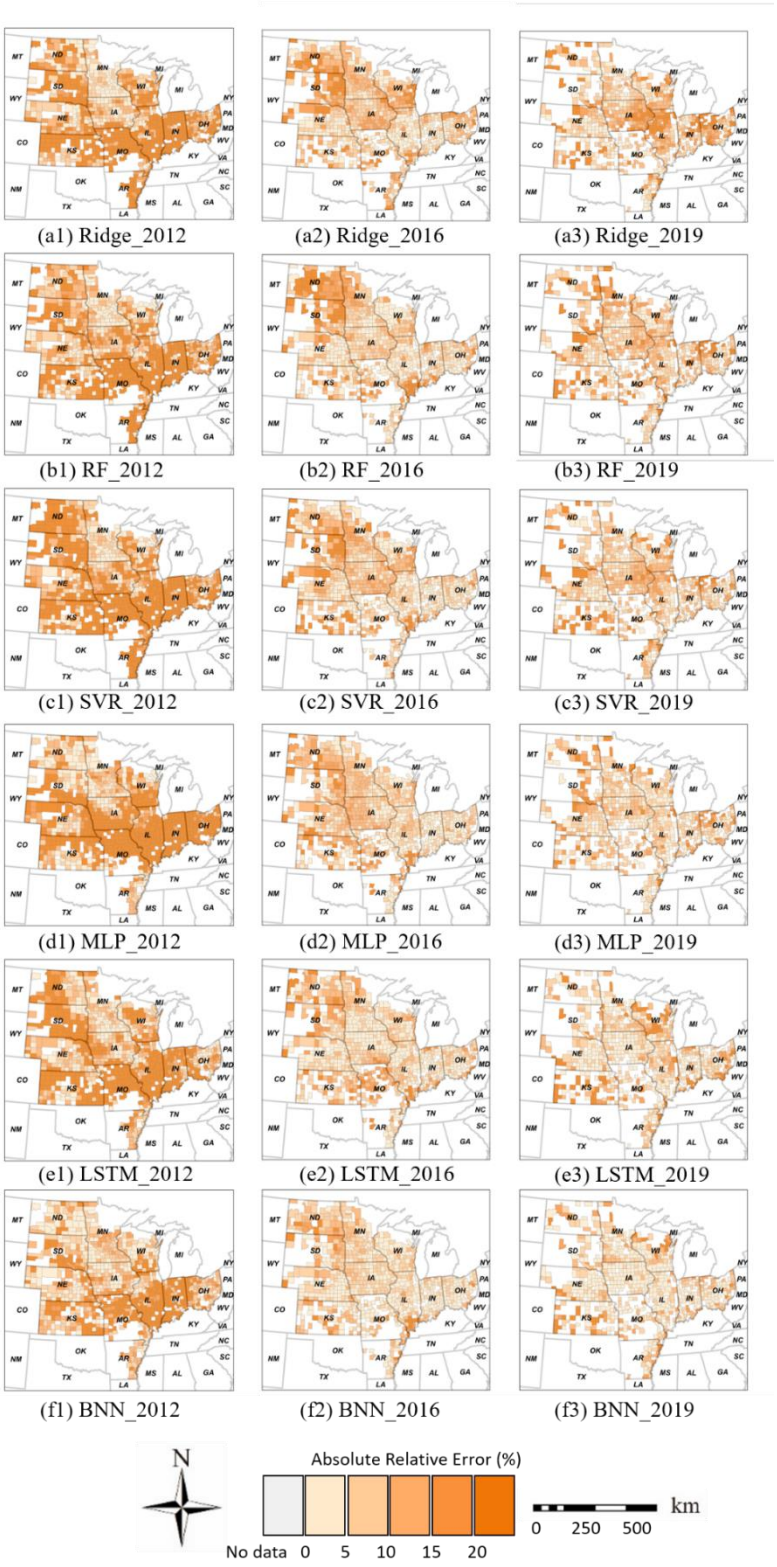

Figure 3-4. The absolute relative error maps of (a) Ridge, (b) RF, (c) SVR, (d) MLP, (e) LSTM, and (f) BNN in three testing years: (1) 2012; (2) 2016; (3) 2019.

3.4.2 In-season Prediction Performance

We further compared the model performance for in-season prediction. Within the growing season from May to October in 2010-2019, we compared the developed BNN model with other non-linear ML methods, and the linear Ridge regression model was excluded from the comparison due to its unsatisfactory performance shown in Section 3.4.1. To present the overall model performance, we averaged the R^2 of each model over ten testing years. The average time series R^2 achieved by each model from middle May to early October at every 16 days in ten testing years are shown in Figure 3-5. In general, it was notable that all the methods performed poorly during the early growing season (before mid-June) when the corn had just been planted or emerged from the ground. During this stage, the RS and weather features had relatively weak correlations with corn yield and it was challenging to make accurate yield predictions (Johnson, 2014). Along with the active growth of corn on the ground, the prediction accuracy gradually increased as more information became available and was captured by the predictors. After that, the model performance became stable in early August when corn transited from the vegetative stage to the reproductive stage. Moreover, compared to the other five approaches, the developed BNN model became the best-performing one since late-June and achieved a near-optimal accuracy in mid-August which is over two months before the harvest season. It demonstrated that the model was able to give highly accurate yield predictions in the middle of the growing season.

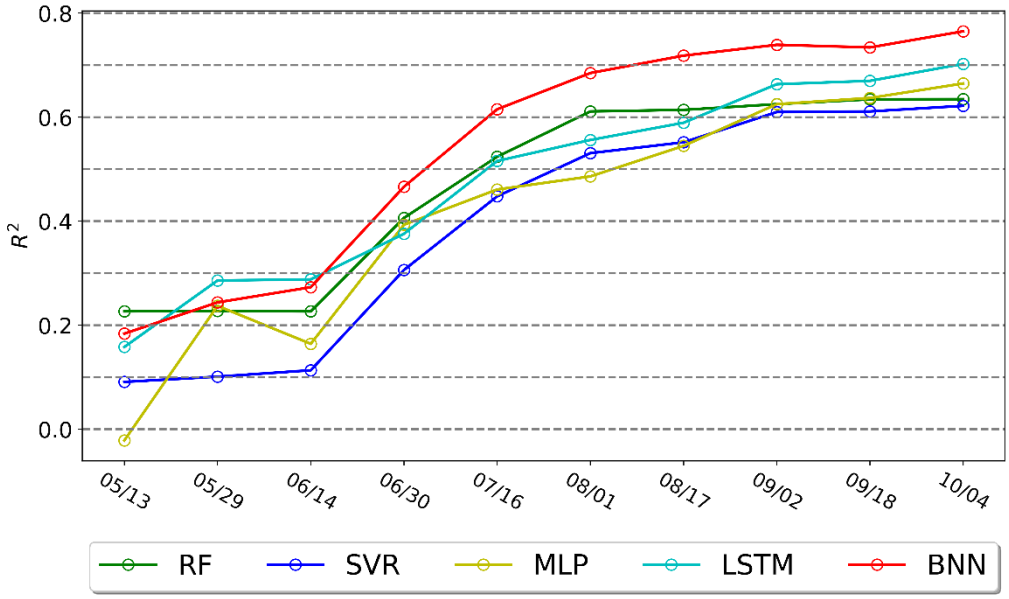

Figure 3-5. The average R^2 of different models during the growing season in all testing years.

Besides the average time-series R^2, we presented the time-series relative error maps of the BNN model averaged over all testing years in Figure 3-6. The figure showed that during the early growing season, large errors were observed in most states. As time proceeded and more informative predictors were used, fewer errors were exhibited and significant improvements were demonstrated in Iowa, Wisconsin, Illinois, and Indiana, where more than half of the counties had an absolute relative error that was less than 10%. Additionally, the spatial error patterns had stabilized since early August, which agreed with the results shown in Figure 3-5. The benefit of more temporal information was less pronounced in North Dakota, South Dakota, Kansas, Missouri, and Arkansas, where crops experienced severer environmental stresses compared to other states (Li et al., 2019; Lobell et al., 2013).

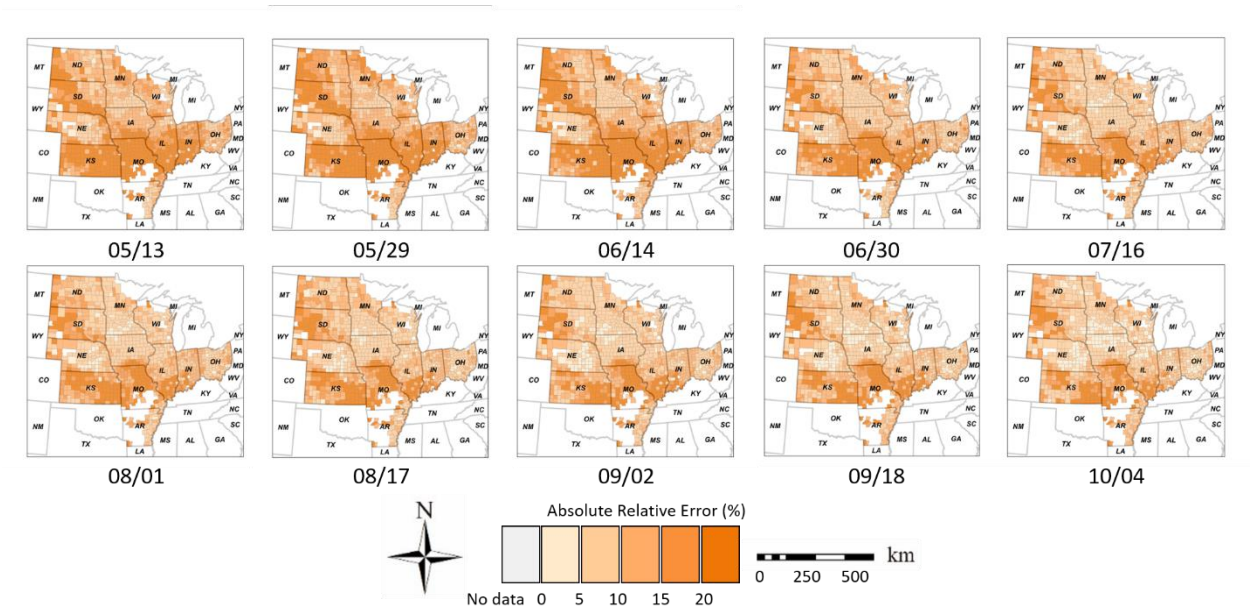

Figure 3-6. The time-series absolute relative error maps for the developed BNN model averaged over all testing years.

3.4.3 Predictive Uncertainty Analysis

Besides the predicted yield, another important output of the developed BNN model was the predictive uncertainty. We first assessed the predictive uncertainty by using the P-factor which is defined as the percentage of observed yield enveloped within the 95% confidence interval bounded by the predictive uncertainty (Sheng et al., 2019) and it can be calculated as below:

$$\text{P-factor} = \frac{NQ_i}{n} \tag{3.8}$$

where NQ_i denotes the number of data samples whose observed yields are within the 95% credible interval of the prediction; n is the total number of data samples. In this study, the P-factor for late-season corn yield prediction across all counties in 2010-2019 was 0.841, showing that more than 84% of observed yield records were successfully enveloped in the confidence interval bounded by the predictive uncertainty. Based on the corn harvest areas in each county, we further grouped all counties into three groups, including the low-area region, the middle-area

region, and the high-area region. Each of them had an equal number of counties. The P-factors for the low-area region, the middle-area region, and the large-area region were 0.851, 0.837, and 0.835, respectively. It demonstrated that the uncertainty estimation worked equally well in counties with different corn harvest areas.

We further analyzed the in-season predictive uncertainty. Here we used the relative predictive uncertainty to represent the normalized uncertainty level in each county, which was defined as a ratio of the predictive uncertainty $\hat{\sigma}$ and the predictive yield $\hat{y}$. The resulting time-series relative uncertainty maps averaged over all testing years 2010-2019 are given in Figure 3-7. The results indicated that predictions in May had the highest uncertainty, reaching more than 10% for most counties. This agrees with the fact that predictions during the early growing season could be unreliable. With more sequential features obtained, the developed BNN model had more confidence in its prediction, and the overall uncertainty level gradually decreased since late May. A stable uncertainty pattern was observed in early August when corn reached the peak of its vegetative stages. It was notable that the pattern of the relative predictive uncertainty (Figure 3-7) mirrored the pattern of the absolute relative errors (Figure 3-6). Mid- and late-season predictions in North Dakota, South Dakota, Kansas, Missouri, and Arkansas had relatively large uncertainties, which is consistent with the prediction errors against the yield statistics provided by NASS. Furthermore, the overall uncertainty change across all the counties within the growing season was summarized in a box plot and shown in Figure 3-8. It was noticed that early-season predictions were associated with much higher uncertainty levels than mid- and late- seasons.

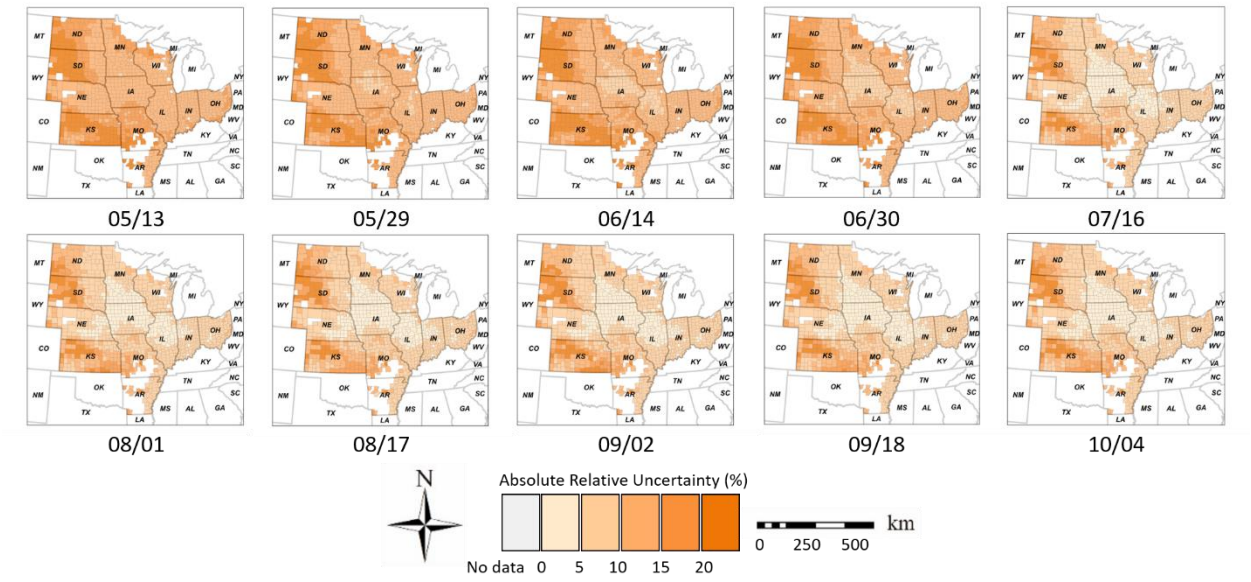

Figure 3-7. The time-series relative predictive uncertainty maps for the developed BNN model averaged over all testing years.

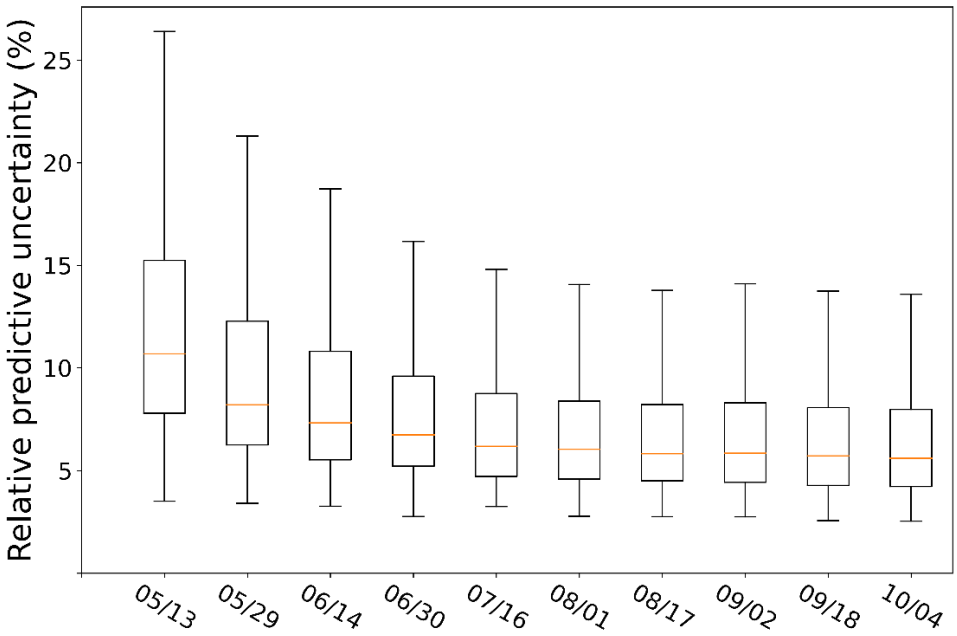

Figure 3-8. Box plot of the time-series relative predictive uncertainty across all the counties.

In general, two main patterns were exhibited in the uncertainty maps (Figure 3-7). On the one hand, as more temporal features were included, the estimations of weight distributions of the BNN model became more accurate, and therefore for all the counties, the uncertainties associated with their predictions were reduced through the growing season. On the other hand, as

we have observed previously, some regions (e.g., North Dakota, South Dakota, Kansas, Missouri, and Arkansas) constantly showed larger prediction errors than others (Figure 3-6) and remained at a relatively high uncertainty level at the end of the growing season (Figure 3-7). This certain spatial pattern was likely caused by the inherent characteristics of the training data. Therefore, we used the results obtained in early September when the uncertainty patterns were stabilized to further explore the causes of this aspect.

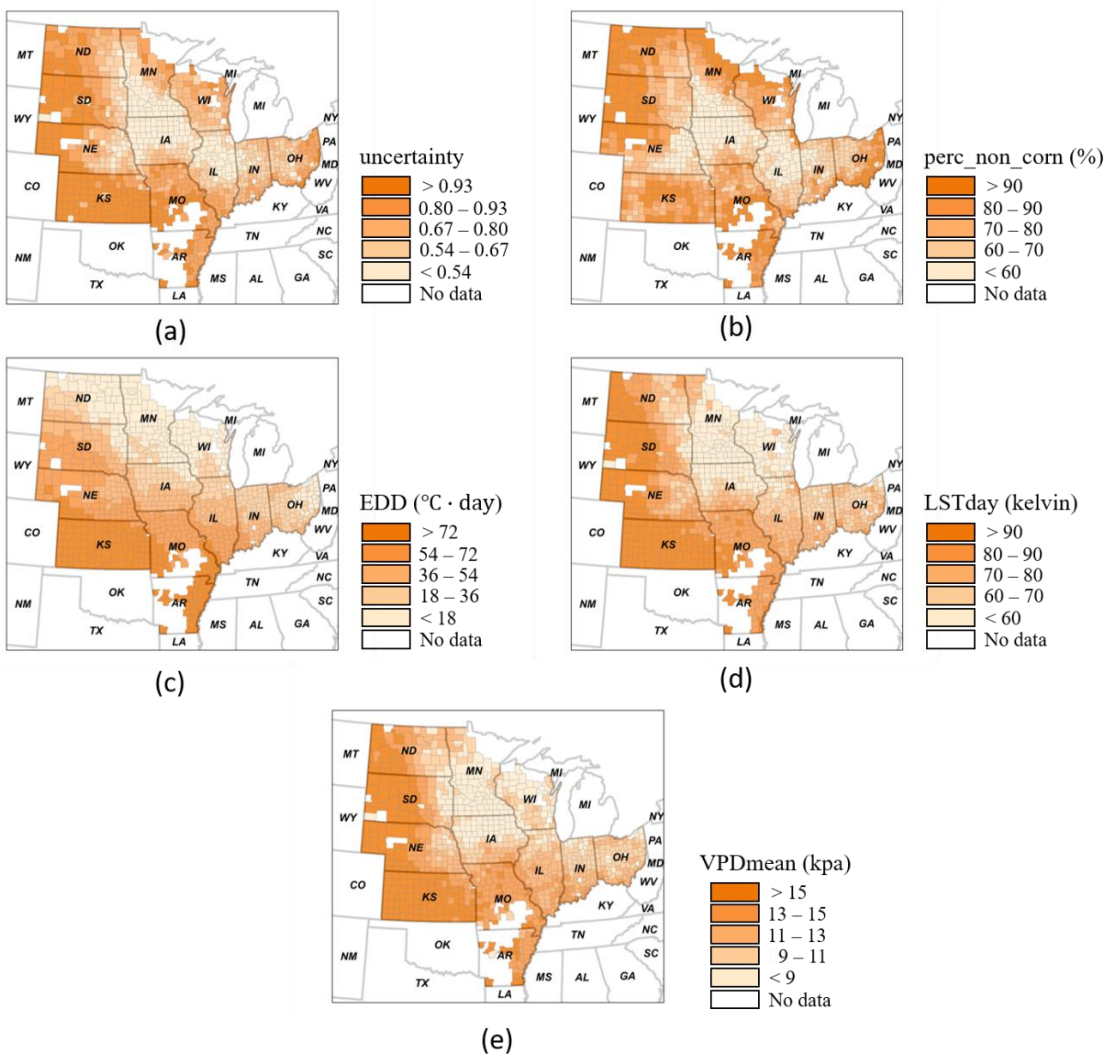

Figure 3-9. Maps of (a) predictive uncertainty on Sep 2nd, (b) percentage of non-corn fields (%), (c) EDD (℃ $\cdot\, day$), (d) LSTday ($Kelvin$), and (e) average VPDmean (kPa).

The observation noise was first considered. In this study, instead of using NASS Cropland Data Layer (CDL) which is not fully available for the U.S. corn belt before 2007, the

MODIS Land Cover Type product was employed as the cropland mask to aggregate the multi-source data in each county for all the years. Since this mask does not distinguish crop types, RS and weather variables on other crops could introduce noise to the training data. We quantified the noise level by calculating the percentage of non-corn fields among all cropland in each county across ten testing years (2010-2019), and the percentage maps (named "perc_non_corn") are shown in Figure 3-9 (b). As expected, a strong similarity pattern was observed between "perc_non_corn" and the corresponding uncertainty maps (Figure 3-9 (a)), with a high positive correlation $r= 0.74$ ($p<0.001$).

Besides the observation noise, we further explored the correlations between the predictive uncertainty with the sequential VI and environmental features, and the results are shown in Figure 3-10. It was noted that lower correlations were shown before the middle of June, and this was mainly because less profound crop growth signals could be captured during the early vegetative stage. Then, during July and August, the predictive uncertainty was found to be strongly correlated with all three VIs. The negative correlations with the VIs indicated that there was less uncertainty for yield modeling in healthier cornfields. Although VIs can be used to indicate crop yield potential, climate variables were the abiotic factors that could significantly affect crop growth and yield potential. In this context, strong correlations were observed between the uncertainty and several weather variables that could potentially relate to two types of environmental stresses: heat stress (Tmax and LSTday) and water stress (VPDmean, VPDmax, ET, and GLDASws). Specifically, during a drought, VPDmean and VPDmax would increase since the actual atmospheric water vapor content is less than the saturated water vapor pressure (Yuan et al., 2019), and ET and GLDASws tend to decrease because transpiration by crops would generally reduce (Kang et al., 2020).

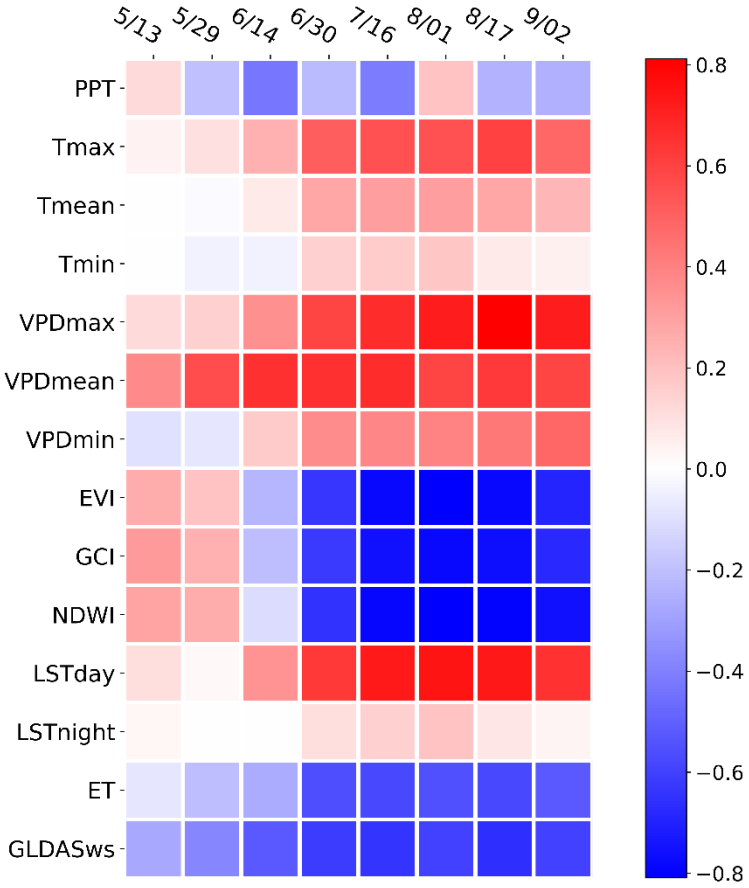

Figure 3-10. Correlation coefficients between time-series features and predicted uncertainty.

However, ET and GLDASws are less representative than VPDmean or VPDmax since evaporation from open water bodies would increase during a drought (Jensen et al., 1990), which resulted in the comparatively weak correlation between the predictive uncertainty with ET and GLDASws. Also, we found that VPDmean and VPDmax were highly correlated. Therefore, we calculated the average VPDmean from June to August to represent the water stress. Also, the Extreme Degree Days (EDD) and average LSTday from June to August were calculated and used as indicators of heat stress that the crop experienced during its critical growth stage. The EDD was considered because it is indicative of cumulative extreme heat within the growing period (Lobell et al., 2013), and it was calculated as below (Eq. 3.9&3.10):

$$EDD = \sum_{t=1}^{N} DD_{30+,\mathrm{t}} \tag{3.9}$$

$$DD_{30+,\mathrm{t}} = \begin{cases} 0 & \text{if } T_{\mathrm{t}} < 30℃ \\ \dfrac{T_{\mathrm{t}} - 30}{24} & \text{if } T_{\mathrm{t}} \geq 30℃ \end{cases} \tag{3.10}$$

in which T_{t} is the hourly temperature estimated based on Tmin and Tmax; $DD_{30+,\mathrm{t}}$ represents the EDD for hour t in each day; N is the total number of hours from June 1st to August 31st.

The resulting maps for the three variables are shown in Figure 3-11 (c)-(e). The results showed that high similarities were observed between the three variables (Figure 3-11 (c)-(e)) with the uncertainty (Figure 3-11 (a)), and the average correlations respectively achieved 0.70 ($p<0.001$), 0.76 ($p<0.001$), and 0.72 ($p<0.001$) for EDD, LSTday, and VPDmean, respectively. Moreover, we noticed that counties with higher predictive uncertainties were mainly located in the western and southern U.S. corn belt where the corn was under heavier heat stress and water stress during the summer (Lobell et al., 2014; Zipper et al., 2016), while in humid and temperate regions (e.g. central U.S. corn belt), lower uncertainty was typically associated with the predictions (Russello, 2018).

3.4.4 Uncertainty Component Analysis

The predictive uncertainty mainly consists of the epistemic uncertainty and the aleatoric uncertainty (Kendall and Gal, 2017). The epistemic uncertainty represents the uncertainty in model parameters and the model structure since we are uncertain about which model parameters or which model structure to choose for prediction. Therefore, we also referred it as the model uncertainty. The epistemic uncertainty can be reduced if more training samples are available, and the model are better trained. On the other hand, the aleatoric uncertainty captures noise inherent in the observations as a result of measurement imprecision. Therefore, we referred it as the data

uncertainty. Since it is caused by the inherent noises of the data, it cannot be reduced even if we have more data.

The predictive uncertainty $\hat{\sigma}$ from the proposed BNN model combines the aleatoric uncertainty $\hat{\sigma}_a$ and epistemic uncertainty $\hat{\sigma}_e$. To quantify the epistemic uncertainty, we can use the Markov sampling. Specifically, given the same input feature vector $\hat{\mathbf{x}}$, the trained BNN model can be run K times and output a series of outputs $\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K\}$. In this case, since we use the same input, the variance of the outputs $\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K\}$ would be mainly caused by the uncertainty in the model parameters, i.e., the epistemic uncertainty. As such, the epistemic uncertainty $\hat{\sigma}_e$ can be quantified as:

$$\hat{\sigma}_e = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\hat{y}_k - \hat{\mu})^2} \tag{3.11}$$

$$\hat{\mu} = \frac{1}{K}\sum_{k=1}^{K}\hat{y}_k \tag{3.12}$$

Based on the sampling results, we calculated the sample standard deviation instead of the population standard deviation. Corresponding, the denominator in Formula (3.11) is $K-1$, which guarantees the unbiased estimation of the standard deviation (Härdle and Simar, 2019). After that. we can quantify the epistemic uncertainty $\hat{\sigma}_a$ by:

$$\hat{\sigma}_a = \sqrt{\hat{\sigma}^2 - \hat{\sigma}_e^2} \tag{3.13}$$

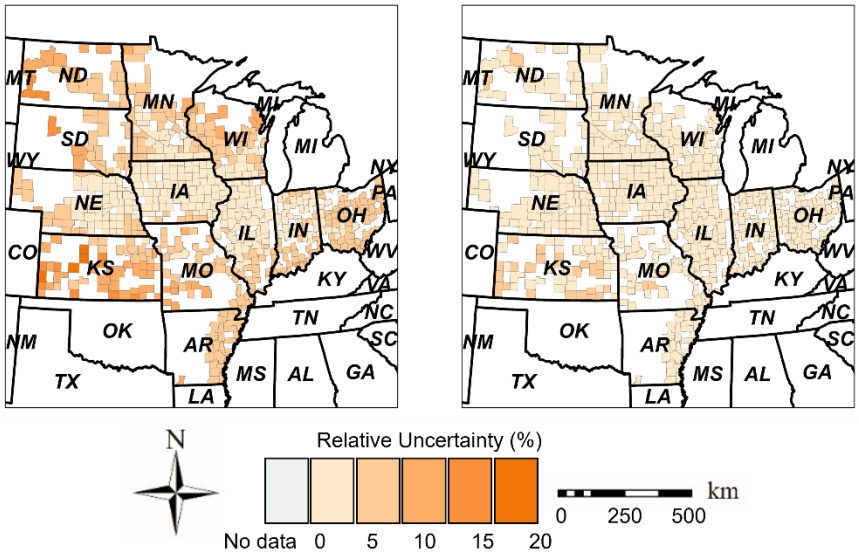

Figure 3-11. The aleatoric uncertainty map (left) and the epistemic uncertainty map (right) in 2019.

We took the testing year 2019 as an example and calculated the aleatoric uncertainty and the epistemic uncertainty for each county. As illustrated in the corresponding relative uncertainty maps (Figure 3-11), the aleatoric uncertainty was larger than the epistemic uncertainty. It demonstrated that the predictive uncertainty was mainly caused by the data noises, which is in agreement with the analysis in Section 3.4.3. On the other hand, since a large number of data had been used for model training, the epistemic uncertainty was reduced to a small level.

3.5 Summary

In this study, we proposed a BNN model for county-level corn yield prediction based on RS images, weather variables, soil properties, harvest year, and historical average yield. The proposed BNN model can simultaneously predict the corn yield and the corresponding predictive uncertainty. Experiments in the U.S. corn belt in ten testing years 2010-2019 showed that the proposed BNN model could make accurate end-of-season yield predictions with stale performances across different testing years. It also outperformed five commonly used linear and non-linear ML and DL models with an average R^2 of 0.77. The in-season prediction performance

was also evaluated within the growing season starting from middle May to early October at a 16-day interval. The developed BNN model achieved near-optimal performance around the middle of August ($R^2 \approx 0.75$), which is about two months ahead of the harvest. Furthermore, we analyzed the time-series predictive uncertainty during the growing season. The results showed that more sequential features could help lower the uncertainty level, and the patterns stabilized around early August. Correlation coefficients analysis showed that the observation noise due to the crop mask, prolonged exposure to extreme heat, and severe water would potentially increase the predictive uncertainty.

The main contributions of this work are summarized as follows:

- A BNN model was proposed for corn yield prediction and uncertainty estimation, which could accurately predict county-level corn yield and outperforms other ML and DL models.
- Accurate corn yield prediction could be made in August, which is about two months ahead of the harvest.
- The predictive uncertainty has a strong correlation with the prediction error.
- The potential sources of predictive uncertainty are observation noises and environmental stresses, such as heat stress and water stress.

CHAPTER 4 ADVERSARIAL UNSUPERVISED DOMAIN ADAPTATION ON CORN YIELD PREDICTION

4.1 Overview

Supervised ML models require data samples with labels (i.e., yield records) for model training. For those agricultural regions without historical yield records, it is impossible to directly train a ML-based crop yield prediction model. Also, due to the domain shift (Kouw and Loog, 2019) caused by spatial heterogeneity of meteorological conditions, soil properties, and farming practice, ML models established between reference (reported) yields and RS measurements within a specific region often lose their validity when directly applied to the other regions. TL, a ML technique that transfers knowledge learned from a local region with rich ground reference data to the target region with limited or no ground truth data, has become a viable solution. To perform TL for deep NN, a widely used strategy is to first pre-train a model on a source domain with abundant ground reference data and then adapt it to a target domain by fine-tuning the pre-trained model with labeled samples from the target domain. However, since collecting yield data can be financially expensive, labor-intensive, and time-consuming, many agricultural production areas may lack reliable ground reference yield data for either directly training or fine-tuning the supervised DL models.

To improve the transferability of DL models without relying on labeled target samples, UDA has been a promising strategy. The core idea of UDA is to reduce the domain shift by aligning the feature distributions in the source domain and the target domain (Figure 4-1). UDA algorithms commonly employ a conjugated architecture with two objectives (Zhao et al., 2020). One objective is to learn a task model based on the labeled source samples by minimizing the

corresponding task loss function, such as MSE for regression (Feng et al., 2021) and cross-entropy loss for classification (Wang et al., 2021). The other objective is to reduce the domain shift and align the source domain and the target domain. One of the most representative single UDA methods is domain adversarial neural networks (DANN) (Ganin et al., 2017), which employs an adversarial objective with a domain discriminator to extract domain-invariant features from source and target domains. The structure of DANN mainly has three parts, including a feature extractor, a domain classifier, and a label predictor. During the training process, the feature extractor is updated to minimize the prediction loss for the label predictor and maximize the domain loss for the domain classier. Since it is trained adversarially against the domain classifier, the feature extractor can be updated towards generating domain-invariant features to help alleviate the negative impact of domain shift. Meanwhile, the feature extractor is trained collaboratively with the label predictor so that it is updated to extract task-informative features to fulfill the main tasks.

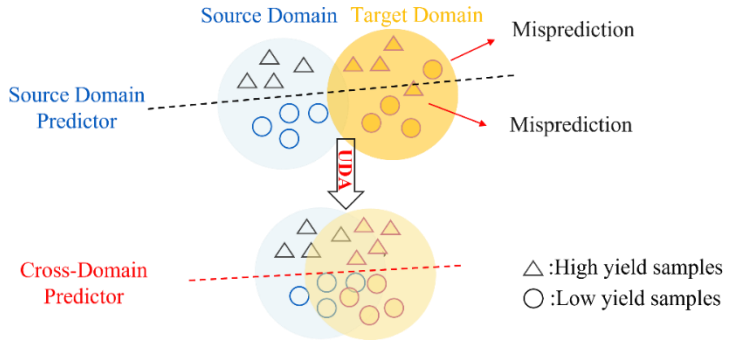

Figure 4-1. A conceptual example of unsupervised domain adaptation (UDA).

Despite the success in classification applications, the DANN model cannot be directly applied for regression tasks such as crop yield prediction. This is mainly because it is hard to find an optimal weighting parameter to control the trade-off between the prediction loss and the domain loss in the original DANN model. Specifically, mean squared error (MSE) is adopted as

the prediction loss which can have a quite different magnitude than the cross-entropy-based domain classification loss. Also, since the training set for county-level corn yield prediction is comparatively small, overfitting may happen during the training of DANN. To address these issues, in **RA-2**, we proposed two variants of DANN, i.e., Adaptive DANN (ADANN) and Bayesian DANN (BDANN), for corn yield prediction at the county level. The ADANN model was designed to adaptively adjust the weighting parameter between the prediction loss and the domain loss to avoid overweighting either of them. Based on ADANN, we further applied Bayesian learning to the model training and designed BDANN intending to improve the model's robustness to overfitting as well as its learning ability on small training sets.

4.2 Methodology

4.2.1 Fundamentals of DANN

Deep NN are trained in a supervised learning way to associate input data samples $\mathbf{x} \in X$ with data labels $\mathbf{y} \in Y$ by learning the distribution $D(\mathbf{x}, \mathbf{y})$, in which X denotes the input feature space and Y denotes the output space. In the scenario of yield prediction, $\mathbf{x}$ are RS and weather variables and $\mathbf{y}$ denotes the reported yield records. Given two datasets from the source domain $\mathcal{D}_s$ with n_s data samples and the target domain $\mathcal{D}_t$ with n_t data samples, they are likely to have different distributions $D_s(\mathbf{x}, \mathbf{y})$ and $D_t(\mathbf{x}, \mathbf{y})$ due to different geophysical environments. Consequently, a ML model trained with data samples from D_s would have degraded performance if directly applied to D_t.

To improve the spatial transferability of DL models across different domains, UDA is a promising strategy. The core idea of UDA is to reduce the domain shift by aligning the feature distributions in the source domain and the target domain. One representative UDA method is the DANN, which reduces the impact of the domain shift by projecting input feature vectors from

two different domains into a common subspace (Ganin et al., 2016; Wang et al., 2018; Zhuang et al., 2019). As illustrated in Figure 4-2, a DANN model mainly consists of three parts, including a feature extractor G_f, a domain classifier G_d, and a yield predictor G_y (Figure 4-2). During training, from left to right, the ith data sample $\mathbf{x}_i$ from $\mathcal{D}_s$ or $\mathcal{D}_t$ is first fed to the feature extractor G_f for feature extraction (Eq (4.1)). Then, the extracted features $\mathbf{x}_i^c$ is forwarded into the domain classifier G_d to predict the domain label $\hat{d}_i$. (Eq (4.2)) The domain label indicates whether the corresponding $\mathbf{x}_i$ is from the source domain ($\mathbf{x}_i \sim D_s(\mathbf{x})$ if $d_i = 0$) or the target domain ($\mathbf{x}_i \sim D_t(\mathbf{x})$ if $d_i = 1$). Meanwhile, the extracted features $\mathbf{x}_i^c$ is forwarded into the yield predictor G_y to predict the yield $\hat{y}_i$.(Eq (4.3)).

$$\mathbf{x}_i^c = G_f(\mathbf{x}_i; \mathbf{w}_f) \tag{4.1}$$

$$\hat{d}_i = G_d(\mathbf{x}_i^c; \mathbf{w}_d) \tag{4.2}$$

$$\hat{y}_i = G_y(\mathbf{x}_i^c; \mathbf{w}_y) \tag{4.3}$$

where $\mathbf{w}_f$ denotes trainable weights in the feature extractor G_f, $\mathbf{w}_d$ denotes the trainable weights in the domain classifier G_d, and $\mathbf{w}_y$ denotes the trainable weights in the yield predictor G_y.

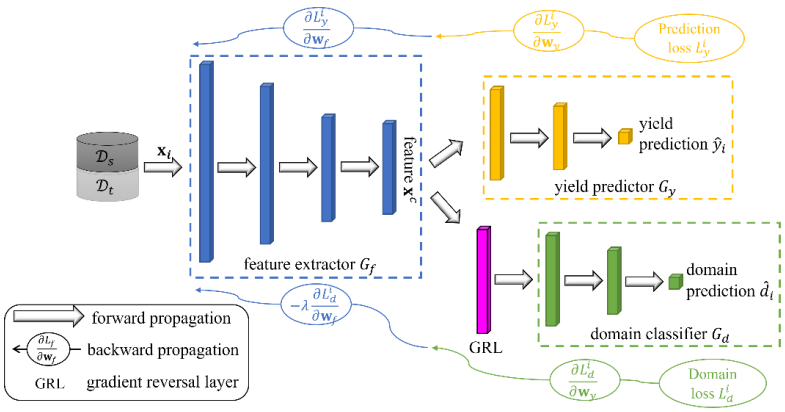

Figure 4-2. The structure of the DANN model, including a gradient reversal layer GRL (pink), a feature extractor G_f (blue), a domain classifier G_d (green), and a yield predictor G_y (yellow).

The model aims to make accurate yield predictions and reduce domain shift through joint training G_f, G_y, and G_d. To achieve these two goals, there are two corresponding training objectives: 1) train the feature extractor G_f collaboratively with the yield predictor G_y to minimize the yield prediction loss L_{yield} and 2) train the feature extractor G_f adversarially against the domain classifier G_d to maximize the domain loss L_{domain}. By realizing the first objective, the feature extractor G_f is updated to extract task-specific features which can be used to accurately predict corn yield by the yield predictor G_y. By realizing the second objective, the feature extractor is driven to project input feature vectors into a common subspace and generate domain-invariant features. To perform adversarial training between G_f and G_d, a gradient reversal layer (GRL) is introduced to connect the domain classifier G_d and the feature extractor G_f. The gradient reversal layer acts as an identity function during forward propagation and reverses the gradient by multiplying it by -1 during backpropagation (Ganin et al., 2017). As a result, the total loss function of this network becomes (Eq. (4.4)):

$$L(\mathbf{w}_f, \mathbf{w}_d, \mathbf{w}_y) = L_{yield} - \lambda L_{domain} \tag{4.4}$$

$$L_{yield} = \frac{1}{n_S}\sum_{i=1}^{n_S}(y_i - \hat{y}_i)^2 \tag{4.5}$$

$$L_{domain} = -\frac{1}{n_S + n_T}\sum_{i=1}^{n_S+n_T} d_i \log(\hat{d}_i) + (1 - d_i)\log(1 - \hat{d}_i) \tag{4.6}$$

where λ is defined as the weighting parameter that adjusts the trade-off between the domain loss and the prediction loss. n_s and n_t represent the number of training samples in the source domain and the target domain, respectively. L_{yield} is the yield prediction loss in the form of MSE (Eq. (4.5)). L_{domain} is the domain loss in the form of binary cross entropy (Eq. (4.6)).

4.2.2 Adaptive DANN

Though successful, the performance of DANN largely depends on the weighting parameter λ between the prediction loss L_{yield} and the domain loss L_{domain}. It is typically predefined experimentally. DANN has been widely used in classification applications where both the prediction loss and the domain loss are defined as the cross-entropy loss and therefore are of a similar magnitude. However, in crop yield prediction, which is a regression task, the yield prediction loss L_{yield} is defined as the form of MSE which can have a quite different magnitude than the cross-entropy-based domain classification loss L_{domain}. Moreover, since the prediction loss L_{yield} can be varying in different regions and different years, it is challenging to find an optimal weighting parameter.

Therefore, we developed an adaptive domain adversarial neural network (ADANN) approach for corn yield prediction (Figure 4-3), in which the weighting parameter λ was adjusted adaptively. Specifically, we followed Ganin et al. (2016) and used a schedule to initialize λ at 0 and gradually increased it as training proceeded. Meanwhile, the λ was normalized with the ratio of the prediction loss and the domain loss to offset the magnitude imbalance. A formal definition of the weighting parameter is given as follows:

$$p_i = \frac{i}{N} \tag{4.7}$$

$$r_i = \frac{L_{yi}}{L_{di}} \tag{4.8}$$

$$\lambda_i = r_i\left(\frac{1}{1+\exp(-p_i)} - 2\right) \tag{4.9}$$

where p_i denotes the learning progress and increases linearly from 0 to 1 as the epoch i increases from 0 to the maximum number of epochs N; r_i is the normalization term defined as the ratio of

the prediction loss and the domain loss during the ith epoch. This schedule makes sure there is a small weight on the domain loss during the early training which enhances the robustness of the domain classifier against noisy signals during the early phase of model training (Ganin et al., 2016). As training proceeds, the weight on the domain loss is increased to make sure that cross-domain features are extracted and can be used for accurate yield prediction by the yield predictors G_y. After convergence, the trained ADANN is able to reduce the domain shift and make accurate yield prediction in the target domain.

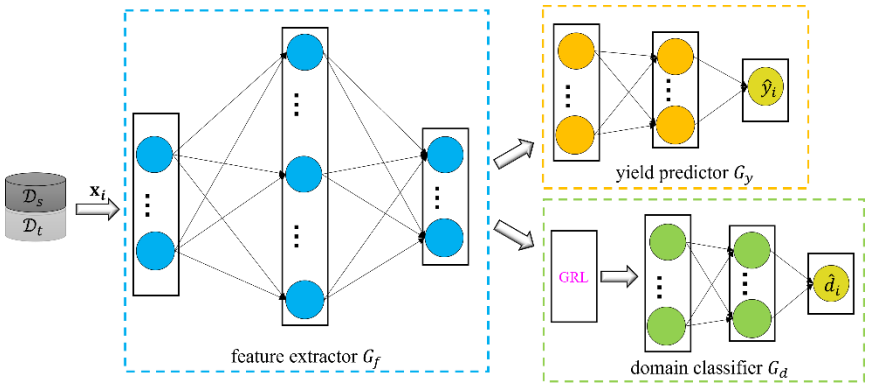

Figure 4-3. The architecture of the ADANN model.

4.2.3 Bayesian DANN

We further proposed the Bayesian domain adversarial neural networks (BDANN) in which we introduced Bayesian inference during model training. There are two major novelties in the proposed BDANN. First, each hidden layer was designed as a Bayesian layer in which the posterior distributions on each weight were approximated via variational inference. By adding the prior data distribution in the model development, BNN is less prone to overfitting due to the prior regularization (Gal et al., 2017; LeCun et al., 2015; Nasrabadi, 2007). Second, to account for the different data uncertainty levels from county to county, the yield predictor was designed to have two endpoints to predict the yield distribution by outputting both the predicted yield value and predictive uncertainty (Figure 4-4).

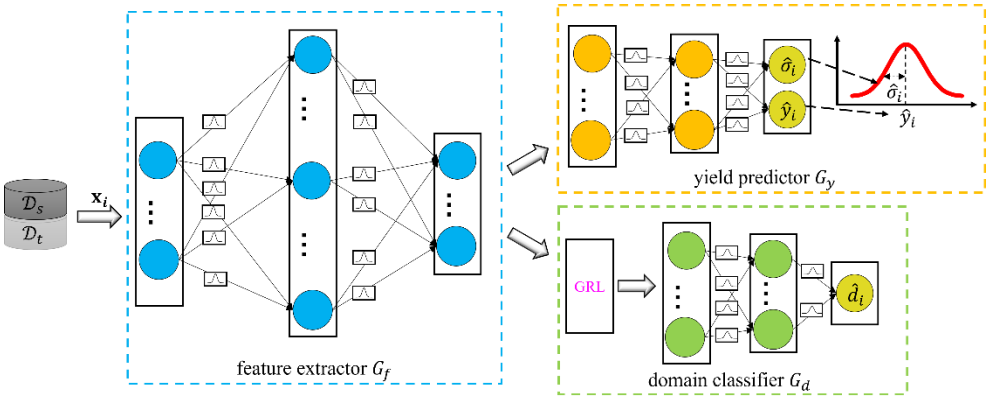

Figure 4-4. The architecture of the proposed BDANN.

Specifically, instead of making point estimates of trainable weights $\mathbf{w}$, BDANN was trained to estimate posterior distributions $p(\mathbf{w}|\mathcal{D})$ given the training data $\boldsymbol{\mathcal{D}}$. However, it is intractable to directly estimate posterior distributions considering the size of deep NN. Therefore, we approximated $p(\mathbf{w}|\mathcal{D})$ by variational learning, which finds the variational distribution $q(\mathbf{w}|\boldsymbol{\theta})$ on the weights given trainable variational parameters $\boldsymbol{\theta}$ that minimizes the *KL*-divergence with the true posterior on the weights (Eq. (4.10)) (Blundell et al., 2015; Ma et al., 2021a):

$$\begin{aligned} L(\boldsymbol{\theta}) &= KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w}|\boldsymbol{\mathcal{D}})] \\ &= \int q(\mathbf{w}|\boldsymbol{\theta}) \log \frac{q(\mathbf{w}|\boldsymbol{\theta})}{p(\mathbf{w})p(\boldsymbol{\mathcal{D}}|\mathbf{w})} d\mathbf{w} \\ &= KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w})] - E_{q(\mathbf{w}|\boldsymbol{\theta})}[\log P(\boldsymbol{\mathcal{D}}|\mathbf{w})] \end{aligned} \tag{4.10}$$

where $KL(p, q)$ denotes the *KL*-divergence between two distributions p and q. There are two terms in (Eq. (4.10)): The first term is a prior-dependent loss $L_{pior} = KL[q(\mathbf{w}|\boldsymbol{\theta})||p(\mathbf{w})]$, which is the *KL*-divergence between the prior and the variational distribution. The second term, $L_{yield} = -E_{q(\mathbf{w}|\boldsymbol{\theta})}[\log P(\boldsymbol{\mathcal{D}}|\mathbf{w})]$, is the negative log-likelihood function for yield prediction. To account for the different data uncertainty levels from county to county, the yield prediction loss L_{yield} was designed in the form of a normal distribution as:

$$L_{yield} = \frac{1}{n_S}\sum_{i=1}^{n_S}\frac{(y_i - \hat{y}_i)^2}{2\hat{\sigma}_i^2} + \log\hat{\sigma}_i \quad (4.11)$$

where $\hat{y}_i$ and $\hat{\sigma}_i$ denotes the mean and the standard deviation for the predictive yield distribution. The mean $\hat{y}_i$ is regarded as the predicted yield and the standard deviation $\hat{\sigma}_i$ quantifies the predictive uncertainty. Note that the yield loss can be converted to MSE (Eq. (4.5)) if $\hat{\sigma}_i$ is assumed to be a constant for all data samples. However, this assumption is invalid in our scenario since the uncertainty level in training data can be varying in different regions as well as in different observation years (Ma et al., 2021a). Therefore, the BDANN was designed to output both the predicted yield and its standard deviation from the predictor, and the yield loss was designed as Eq. (4.11). The final loss function for the proposed BDANN is (Eq. (4.12)):

$$L(\boldsymbol{\theta}) = L_{prior} + L_{yield} + \lambda L_{domain} \quad (4.12)$$

4.3 Experimental Setup

Both ADANN and BDANN were trained and evaluated within twelve Midwestern U.S. states, including North Dakota, South Dakota, Kansas, Nebraska, Minnesota, Iowa, Wisconsin, Illinois, Indiana, Ohio, Missouri, and Michigan. To evaluate the model's transferability across different regions, the counties under study were grouped into two diverse ecological regions based on the United States Environmental Protection Agency (EPA) (EPA, 2001), including the Eastern Temperate Forests (ETF) region and the Great Plains (GP) region (Figure 4-5). The ETF is characterized by a warm, humid, and temperate climate, with humid summers and mild to cold winters. ETF is mainly covered by dense and diverse forests. GP, on the other hand, mainly consists of flat grasslands and has a scarcity of forests. The change of seasons in GP is more obvious, with very hot summers and harsh winters. GP is also subjected to drought, due to the

scarcity of forests and lack of rainfall (Omernik, 1987; Omernik and Griffith, 2014). Therefore, we chose ETF and GP as two domains for transfer experiments.

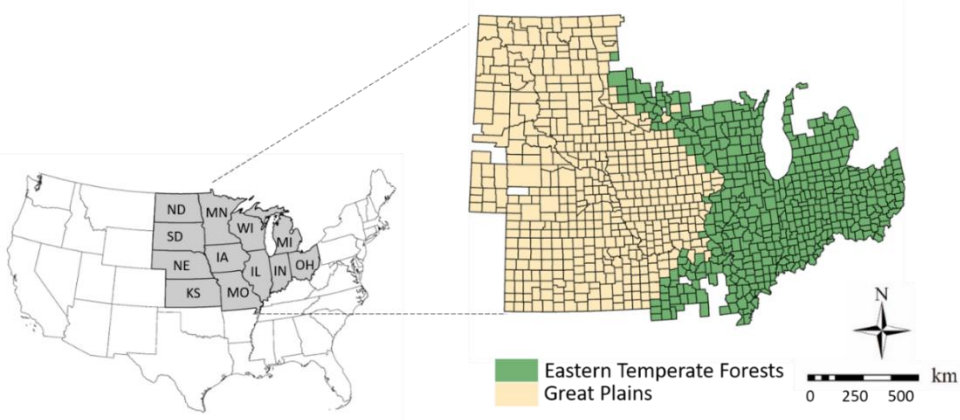

Figure 4-5. Corn growing counties in the two ecosystem regions within the study area, including the Eastern Temperate Forests (ETF) region and the Great Plains (GP) region.

Time-series RS imagery and weather observations were collected from 2006 to 2019 as input predictors (Table 4.1). Three representative VIs (i.e., EVI, NDWI, and GCI) and five weather variables (LSTday, LSTnigh, Tmax, Tmean, and PPT) were selected as the predictors and paired with the reported yield records from USDA for model development (USDA, 2020b).

Table 4-1. Summary of study areas and data used for model development.

Domains	Environment and Climate	Landcover Layer	Predictor Variables
Eastern Temperate Forest (ETF)	Largely covered by dense forests with humid summers and temperate winters	USDA-NASS Cropland Data Layer (CDL)	• EVI, NDWI, GCI from MODIS MCD43A4 • LSTday and LSTnight from MODIS MYD11A2 • Tmax, Tmean, PPT from PRISM
Great Plains (GP)	Mainly consists of grasslands and a scarcity of forests, with very hot summers and harsh winters.		

Besides ADANN and BDANN, three other approaches were chosen as the comparison methods, including RF, DNN, and the original DANN model with a fixed weighting parameter λ in the loss function. Each model was evaluated under two transfer experiments (i.e., GP→ETF and ETF→GP) where GP and ETF were alternatively used as the source domain and the target domain for model evaluation. For RF and DNN, they were trained only using labeled samples from the source domain and directly evaluated in the target domain. For the UDA method, they were trained using labeled source samples and unlabeled target samples and evaluated in the target domain in each testing year. Using all preceding years since 2006 for model training, the models were evaluated in four testing years 2016 - 2019. R^2 and RMSE were selected as the metrics to evaluate the performance of each model (Eq. (3.5-3.6)).

4.4 Results and Discussion

4.4.1 Experimental Results

The evaluation results of transfer experiments from GP to ETF and from ETF to GP were shown in Table 4.2. The best-performing one was highlighted in bold for each study case. It was observed that both RF and DNN had poor performances in the target domain. Due to the domain shift existing between the source domain and the target domain, RF and DNN trained in the source domain could not accurately predict the corn yield in the target domain. Especially in 2019, when a flood postponed the planting of corn in several states in ETF (i.e., Illinois and Indiana) (Baum et al., 2020), the domain shift was further enlarged between ETF and GP, and both RF and DNN achieved low agreement in the transfer experiment from GP to ETF.

Through UDA, it was observed that the DANN model had improved performance in the target domain in several cases. For example, in the year 2019, DANN improved the R^2 and

RMSE to 0.42 and 1.08 t/ha in the transfer experiment from GP to ETF, respectively. However, DANN had an unstable performance and could only slightly improve the accuracy in several transfer experiments when compared to RF and DNN. ADANN and BDANN further outperformed the DANN with more stable performance in different testing years. ADANN and BDANN had similar performance in most cases while BDANN outperformed ADANN in several cases. For example, when transferring from GP to ETF in 2019, BDANN decreased the RMSE by about 10% when compared to the ADANN model. It demonstrated that by estimating the weights in the form of variational distributions, BDANN was more robust and had stronger learning abilities.

Table 4-2. Model evaluation in transfer experiments.

Year	Case	RF		DNN		DANN		ADANN		BDANN	
		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
2016	GP→ETF	1.26	0.49	1.14	0.58	1.12	0.60	0.96	0.70	**0.90**	**0.74**
	ETF→GP	1.70	0.28	1.63	0.33	1.50	0.44	1.16	0.66	**0.96**	**0.76**
2017	GP→ETF	1.01	0.56	1.15	0.45	0.88	0.58	0.84	0.60	**0.83**	**0.61**
	ETF→GP	1.83	0.47	1.45	0.67	1.38	0.70	**1.06**	**0.82**	1.20	0.77
2018	GP→ETF	1.10	0.57	1.22	0.47	1.17	0.51	0.98	0.66	**0.95**	**0.67**
	ETF→GP	1.68	0.45	1.75	0.41	1.42	0.61	1.40	0.62	**1.18**	**0.72**
2019	GP→ETF	1.23	0.26	1.27	0.20	0.96	0.48	0.93	0.51	**0.88**	**0.57**
	ETF→GP	1.43	0.59	1.21	0.72	1.13	0.76	**1.07**	**0.78**	**1.07**	**0.78**

To evaluate whether the methods are statistically different on the reported R^2, the paired sample t-test between the ADANN and BDANN evaluation results and each comparison model were performed. A t-test is a statistical test that compares the means of two samples (Mishra et al., 2019). In our case, we first compared the means of R^2 of ADANN model with each model in all testing years. Then, we compared the means of R^2 of BDANN model with each model in all

testing years. Since each experiment was repeated five times, there were a total of 40 pairs of samples in the t-test. As shown in Table 4-3, the results showed that the accuracy improvement obtained by the proposed ADANN and BDANN was statistically significant.

Table 4-3. Results of the paired sample t-test between the R^2 of each model in all testing years.

Model	t	*p*-value
ADANN vs RF	39.445	0.000
ADANN vs. DNN	28.750	0.000
ADANN vs. DANN	17.434	0.000
BDANN vs RF	49.461	0.000
BDANN vs. DNN	35.548	0.000
BDANN vs. DANN	26.554	0.000
BDANN vs ADANN	8.083	0.000

We further draw the density scatter plots of reported corn yield versus predicted corn yield in all testing years 2016-2019 to show the agreement of reported yield and predicted yield (Figure 4-6). It was observed that both RF and DNN were unable to achieve good agreement in both transfer experiments. Specifically, in the transfer experiment GP → ETF, severe underestimation happened since most of the scatter points were located below the reference line (Figure 4-6 (a1)-(b1)). In the transfer experiment ETF → GP, underestimation still happened to high-yielding counties in GP (Figure 4-6 (a2)-(b2)). The main reason was that the domain shift would cause bias in model training and result in biased prediction when directly applying the model to the target domain. This biased estimation was significantly mitigated by DANN when UDA was considered. However, DANN still had biased estimations and overestimated corn yield for low-yielding counties in ETF (Figure 4-6 (c1)). ADANN and BDANN further outperformed DANN and achieved better agreement between the reported yields and the predicted yields. It demonstrated that ADANN and BDANN had better spatial transferability (Figure 4-6 (d)&(e)).

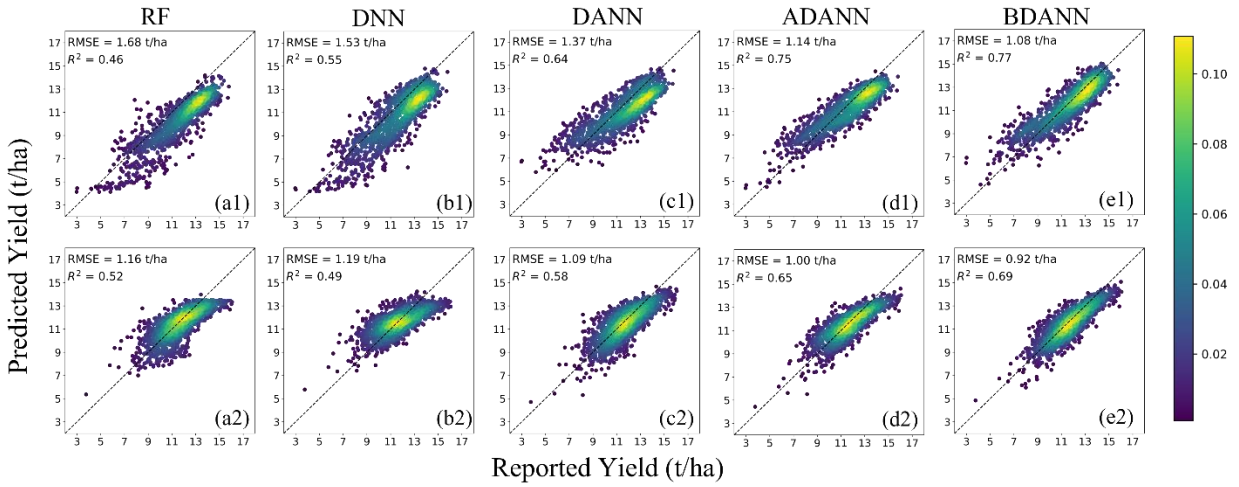

Figure 4-6. The density scatter plots of reported corn yield versus predicted corn yield in all testing years 2016-2019 in transfer experiments (1) ETF → GP and (2) GP → ETF for model (a) RF, (b) DNN, (c) DANN, (d) ADANN, (e) BDANN

Moreover, we present the absolute error maps averaged over years 2016-2019 for each model (Figure 4-7). For RF and DNN, clusters of large errors were observed in regions far from the source domain. For example, when transferring from ETF to GP, a large number of errors were observed in the west of Nebraska and Kansas (Figure 4-7 (a1)-(b1)). Similarly, when transferring from GP to ETF, large errors were concentrated in the east of Michigan and Ohio (Figure 4-7 (a2)-(b2)). Without UDA, RF and DNN tended to make large errors in these distant areas that have large domain shifts with the source domain. The DANN model had reduced the errors but still had clusters of large errors in the west of Nebraska (Figure 4-7 (a3)). ADANN and BDANN further reduced the errors and had better spatial transferability across the target domains (Figure 4-6 (d)&(e)).

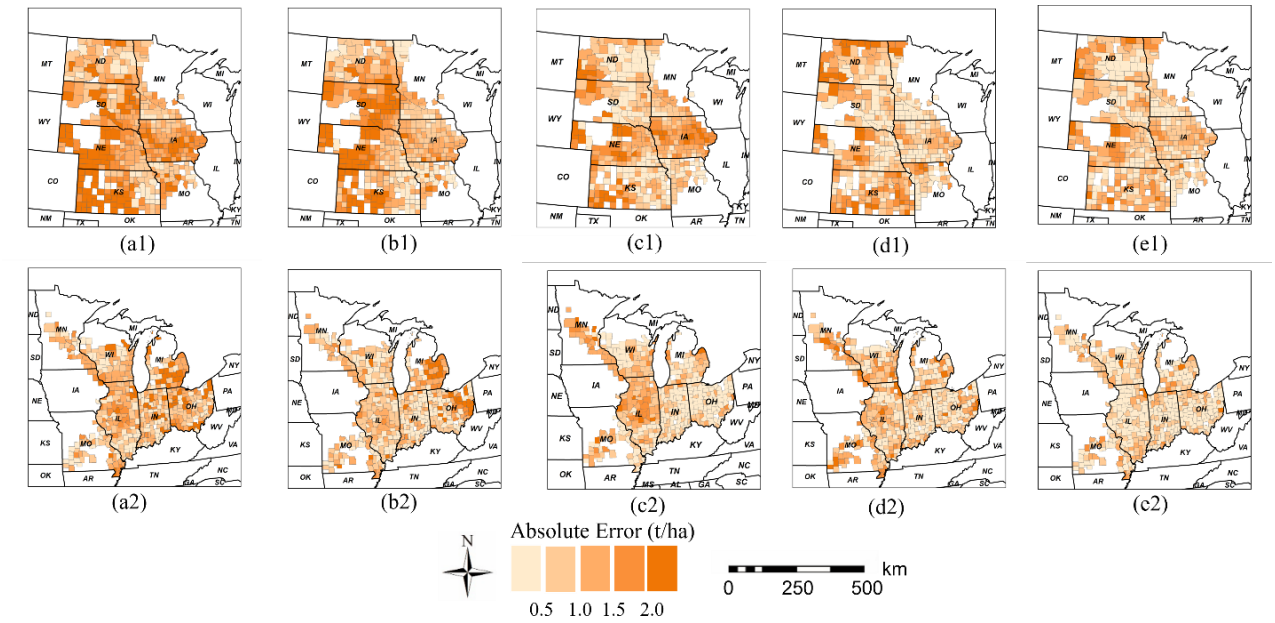

Figure 4-7. The absolute error maps averaged over years 2016-2019 in transfer experiments (1) ETF → GP and (2) GP → ETF for model (a) RF, (b) DNN, (c) DANN, (d) ADANN, (e) BDANN.

4.4.2 t-SNE Visualization of Feature Distributions

To provide a visual insight into the effects of UDA by each model, we visualized the feature distributions of the input feature vectors as well as the extracted features by DANN, ADANN, and BDANN using the t-distributed Stochastic Embedding (t-SNE) algorithm. The t-SNE algorithm is an unsupervised data visualization tool that projects the high-dimensional feature vectors to a low-dimensional space (i.e., two-dimension in our case) for visualization (Maaten and Hinton, 2008). Since it is challenging to do spatial analysis in the original high-dimensional feature space, the t-SNE algorithm was used to project a high-dimensional feature vector to a two-dimensional space for visualization. Figure 4-8 shows the visualization results for the original input features and extracted features by each UDA model in 2019. Data samples from ETF and GP are color-coded, with green representing ETF and yellow representing GP.

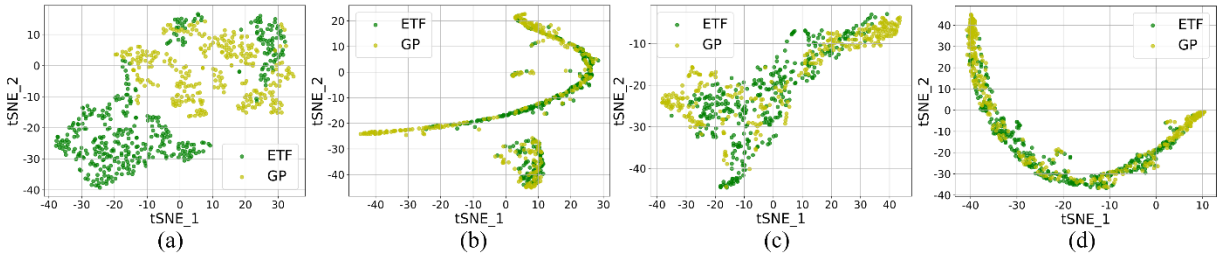

Figure 4-8. The t-SNE visualization results of (a) original features and extracted features by (b) DANN, (c) ADANN, and (d) BDANN in 2019.

Before UDA, the original feature distributions from ETF and GP were separate from each other with limited overlapping areas (Figure 4-8 (a)). It indicated that there existed large domain shifts between ETF and GP. Therefore, RF and DNN trained in one domain would have poor performance in the other domain. The goal of UDA is to reduce the domain shift between source and target domains by blending features from different domains into a uniform distribution in the feature space. After UDA, DANN, ADANN, and BDANN were able to largely reduce the domain shift (Figure 4-8 (b)-(d)). However, some isolated data samples were observed in the visualization results of DANN (Figure 4-8 (b)). It was found that those isolated data samples were mostly counties near the boundary of two domains, such as from Wisconsin and Minnesota. Since those counties were geographically close to each other and tended to have small domain shifts, the domain loss calculated based on those data samples was always small. Therefore, nearby data samples could be easily aligned with each other. Consequently, the feature extractor of DANN would be mostly updated based on distant data samples with large domain shifts. Also, since a fixed weighting parameter was given between the yield prediction loss and the domain loss, DANN was trained to reduce the domain loss from the beginning of training. As a result, DANN was updated to separately match nearby data samples and distant data samples but failed to consider aligning the whole source and target domains (Ma and Zhang, 2022).

On the other hand, ADANN was able to adaptatively adjust the weighting parameter. At the beginning of training, the domain loss was weighted near zero so that the model would not be mostly updated based on distant data samples with large domain shifts. As the training proceeded, the weighting parameter increased and ADANN gradually align the source domain and the target domain. Therefore, the data samples from both domains were well aligned. Moreover, BDANN, due to the regularization by the prior distributions, could further prevent being overtrained on distant data samples and thus learned to extract cross-domain features with high spatial generalizability. Therefore, the t-SNE results showed that ADANN and BDANN had closely aligned the whole source and target domains with no isolated data clusters (Figure 4-8 (c)-(d)).

4.4.3 Model Performance with Different Sizes of Training Sets

Finally, the learning abilities of ADANN and BDANN on small training datasets and their performances were evaluated under different sizes of the training set. Specifically, the training set with data samples from 2006 to 2018 was decreased and used to train the model. After that, the trained model was evaluated on the full testing set in the target domain in the testing year 2019. ETF and GP were alternatively used as the source domain and the target domain. In both experiments, the size of the training set was gradually decreased from 100% (the whole training set) to 10% (only 10% of the training set was kept). The model performance was shown in Figure. 4-9.

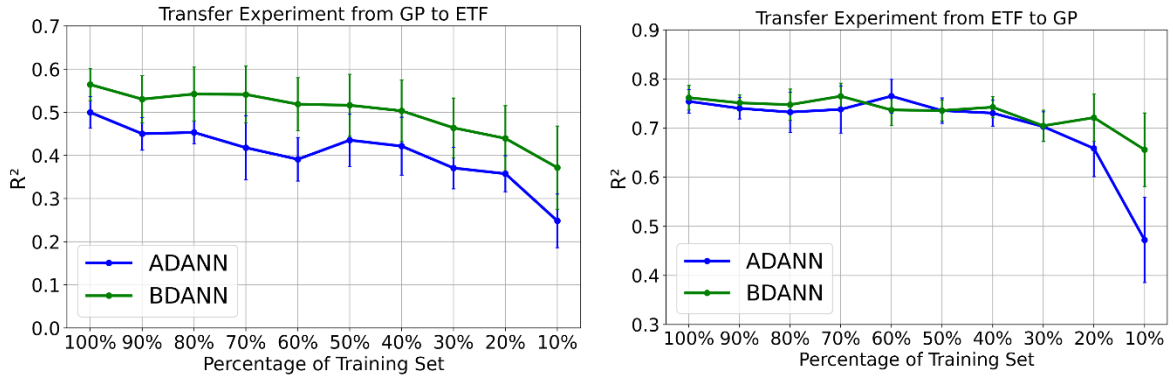

Figure 4-9. Mean R^2 and its standard deviation of ADANN and BDANN in transfer experiments (1) GP → ETF and (2) ETF → GP when reducing the training set size from 100% to 10% in 2019.

It was observed that the BDANN model had a more stable performance in both transfer experiments (Figure 4-9). Specifically, when transferring from GP to ETF, BDANN constantly outperformed ADANN regardless of the training size. Also, the ADANN model's performance decreased significantly when only a small percentage of the training set was used. On the other hand, when transferring from ETF to GP, both ADANN and BDANN performed equally well when a relatively large training set was used. This was because the training set in the ETF was more equally distributed. Therefore, when only a part of the training set in ETF was left, both ADANN and BDANN could still well align ETF and GP. As the training size was further decreased, the remaining training set in ETF was less representative and thus more difference between the two models was observed. The BDANN model was able to achieve an R^2 of over 0.60 when using only 10% of training data, while the performance of ADANN dropped significantly. These results demonstrated that the BDANN model could generalize well on small training sets.

4.5 Summary

In this study, the strategy of adversarial domain adaptation was used to improve the spatial transferability of DL models for corn yield prediction. Two variants of DANN, i.e., ADANN and BDANN, were proposed for corn yield prediction at the county level based on RS images and weather variables. The ADANN model was designed to adaptively adjust the weighting parameter between the yield prediction loss and the domain classification loss. Based on ADANN, we further proposed BDANN by applying Bayesian inference to the model training. Both ADANN and BDANN were evaluated in two ecoregions of the U.S. corn belt and compared with other widely used ML models and UDA methods including RF, DNN, and the DANN model with a fixed weighting parameter. Evaluation results across two ecoregions in the U.S. corn belt demonstrated that the proposed ADANN and BDANN had better spatial transferability with more stable performance against RF, DNN, and DANN in four testing years 2016-2019. The t-SNE visualization results showed ADANN and BDANN could effectively reduce the domain shift and well align the source domain and the target domain. Furthermore, we also compared the model performance on the training set with decreasing size. It was observed that the BDANN model could generalize well on small training sets.

The main contributions of this work are summarized as follows:

- The UDA strategy was used for corn yield prediction and two adversarial domain adaptation models were proposed for county-level corn yield prediction based on RS images and weather variables.
- The proposed ADANN and BDANN outperformed RF, DNN, and DANN with better spatial transferability across spatial domains.

- The t-SNE visualization showed that ADANN and BDANN were able to effectively reduce the domain shift and well align the source domain and the target domain.
- Experiments on the training set with a decreasing size demonstrated that the BDANN model could generalize well on small training sets.

CHAPTER 5 MULTI-SOURCE MAXIMUM PREDICTOR DISCREPANCY FOR UNSUPERVISED DOMAIN ADAPTATION ON CORN YIELD PREDICTION

5.1 Overview

In Chapter 4, we have demonstrated the effectiveness of the UDA strategy on corn yield prediction. However, there are two major bottlenecks in applying current UDA methods to crop yield prediction based on RS images and weather variables.

First, most existing UDA methods attempt to address the domain shift by directly matching feature distributions in the source and target domains by extracting domain-invariant features. However, directly aligning feature distributions may cause ambiguous domain-invariant features. Such ambiguous features have limited information regarding the main task (i.e., crop yield prediction). For example, in the scenario of crop yield prediction, large domain shifts are likely to exist across geospatial domains due to environmental variations (Deines et al., 2021). For those target samples that are outside the support of the source domain, they are likely to lose discriminative features if the crop yield response in the target domain is not considered during UDA. Therefore, simply aligning the feature distributions without considering the specific task for the target domain may not effectively address the domain shift and the prediction accuracy in the target domain could be still low after UDA (Riemer et al., 2019).

Second, most UDA methods were designed for single-source UDA, which assumes that all the labeled data are collected from a single homogeneous domain. In reality, crop yield data used for model training are typically collected from multiple heterogeneous regions (Zhao et al.,

2020). In such cases, the single-source UDA strategy could be trivially applied by combining the different regions into a single source (Ma et al., 2021b). However, due to the spatial heterogeneity, the domain shift not only exists between the source domain and the target domain but also exists among different source domains/regions. Thus, the combined data across multiple regions may negatively interfere with each other during the learning process.

A promising way to address the first issue is to use the Maximum Classifier Discrepancy (MCD) instead of using a domain classifier/discriminator during adversarial learning (Saito et al., 2018b). The core idea of MCD is to align source and target distributions by utilizing the task-specific decision boundaries modeled by two classifiers. First proposed for image classification by Saito et al. (2018b), Satio et al. trained a feature extractor and two independent classifiers on labeled source images and unlabeled target images. The two classifiers are trained to minimize the classification loss on labeled source images while maximizing their classification discrepancy on unlabeled target images. By doing this, the model has been trained to measure the domain shift in a task-specific way. The model could distinguish target images that are far from the support of the source domain. The feature extractor is then trained to fool the classifiers by minimizing the classifier discrepancy to have the target samples generated inside the support of the source which can help reduce the domain shift.

To address the second issue and better leverage data from multiple source domains, there have been growing interests in multi-source UDA, which is a powerful extension of single-source UDA that aligns the target domain to multiple source domains simultaneously (Zhao et al., 2020). Recent multi-source UDA models are mostly developed by extending existing single-source UDA strategies. For example, Peng et al. (Peng et al., 2019) extended the single-source moment-matching method and proposed a moment-matching multi-source UDA model named

M^3SDA for image classification, which reduces source-target divergences and inter-source divergences by minimizing the moment-related distances between each domain. Similarly, adversarial-based single-source UDA models have been extended for multi-source UDA by incorporating multiple feature extractors and domain discriminators for multi-way adversarial learning. For instance, Xu et al. (Xu et al., 2018) proposed a deep cocktail network that uses multi-way adversarial learning to minimize the discrepancy between the target and source domains for image classification. Zhao et al. (Zhao et al., 2019) designed separate feature extractors for each source to learn discriminative target representations in an adversarial manner for image classification.

Motivated by MCD and the recent development of multi-source UDA, in this chapter, we proposed a novel MMPD model for corn yield prediction using satellite images and weather variables. First, inspired by MCD, we proposed MPD and designed a feature extractor and source-specific yield predictors. The feature extractor and each pair of source-specific yield predictors were trained in an adversarial manner to align the source and target domains by considering crop yield response in the target domain through the yield prediction regression curves. Second, to avoid negative interference among labeled data from heterogeneous spatial regions, the strategy of multi-source UDA was employed by grouping labeled data based on multiple sources and adapting them to the target domain separately. The final predictions on the target domains were made based on the ensemble results from multiple source domains.

5.2 Methodology

5.2.1 Multi-source Maximum Predictor Discrepancy

In the scenario of crop yield prediction, input predictors $\mathbf{x} \in X$ are RS images and weather variables, and the target variable $y \in Y$ is the crop yield. X denotes the input feature

space and Y denotes the label space. Given an unlabeled target domain $\mathcal{D}_t$ and M labeled source domains $\mathcal{D}_s = \{\mathcal{D}_1, \dots, \mathcal{D}_M\}$, the MMPD model has a weight-shared feature extractor G_f, which takes input $\mathbf{x}_i$ from source or target domains to extract features, and M pairs of source-specific yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$, which takes extracted features from G_f and make yield prediction $\hat{y}_{i_k}$ and $\hat{y}'_{i_k}$ (Figure 5-1). G_{p_k} and G'_{p_k} are the pair of yield predictors for the i-th source domain and they have the same structure.

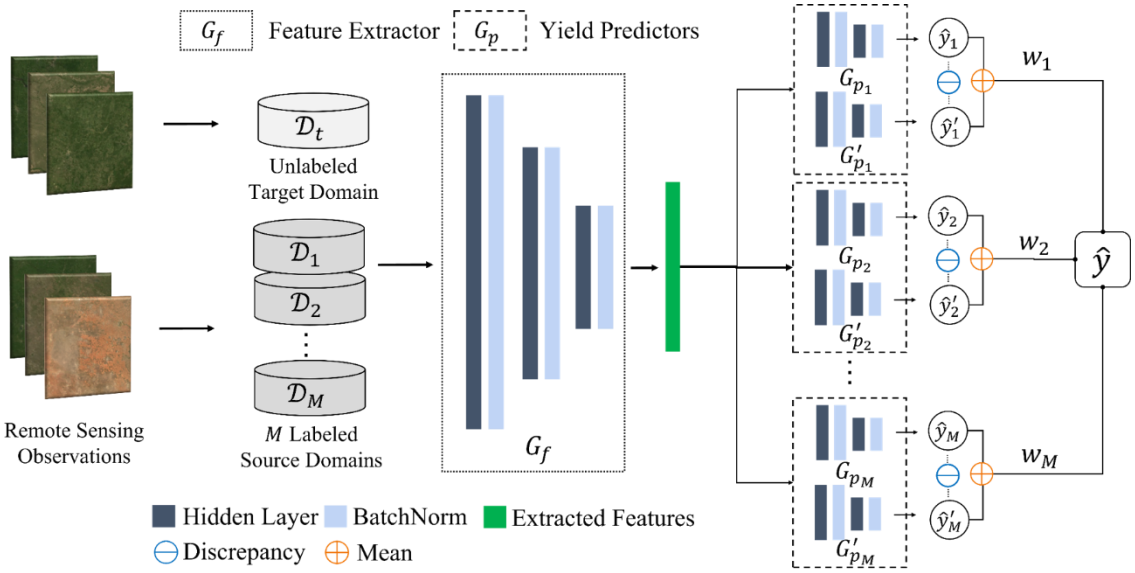

Figure 5-1. The architecture of the proposed MMPD model.

The goal of the MMPD model is to align source and target domains by utilizing a pair of source-specific yield predictors as the discriminator to consider the relationship between regression curves and target samples. For this objective, we need to detect target samples far from the support of each source and align them to the source. Such target samples are likely to be inaccurately predicted by the predictors trained with labeled source samples because they have a large domain shift with the source domain. Therefore, we proposed to utilize the disagreement of the pair of domain-specific yield predictors in yield prediction for target samples. Each pair of domain-specific yield predictors are first trained with labeled source samples so that they can

classify source samples correctly. Note that each pair of yield predictors are initialized with different weights to obtain different predictors at the beginning of training. As a result, for target samples far from the support of the source domains, the pair of source-specific yield predictors are likely to make very different predictions. The disagreement of each pair of yield predictors on target samples, which is termed as predictor discrepancy, can indicate the domain shift between the source and target domains. Each pair of domain-specific yield predictors is trained to maximize the predictor discrepancy on target samples to effectively detect the target samples outside the support of the source domain. On the other hand, the G_f is trained to fool the discriminator by minimizing the predictor discrepancy to have the target samples generated inside the support of the source which can help reduce the domain shift. The goal is to obtain a G_f that can extract domain-invariant and task-informative features. To achieve that goal, the MMPD model is trained recursively in three steps (Figure 5-2).

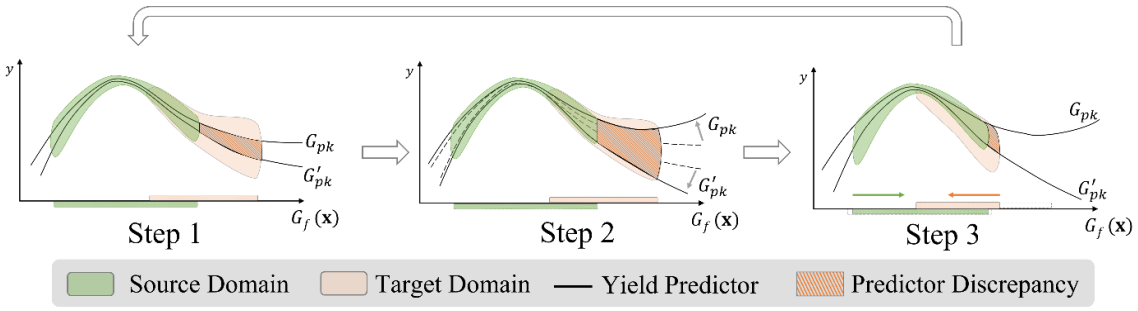

Figure 5-2. The three steps of training the MMPD model.

Step 1: First, to make the weight-shared feature extractor G_f and the domain-specific yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ obtain informative features, they are trained to correctly predict crop yields of the source samples (Figure 5-2. (Step 1)). Specifically, labeled data from each source is forwarded through the feature extractor to extract features. Then, the extracted features are fed forward to the source-specific predictors for yield prediction. Each pair of yield

predictors is trained to be the expert in the specific source domain. The yield prediction loss in each source is calculated as the mean squared error (Eq. (5.2)-(5.3)) and the model is trained to minimize the total yield prediction loss in all sources (Eq. (5.1)):

$$\min_{G_f,\{G_{p_k},G'_{p_k}\}_{k=1}^{M}} \sum_{k=1}^{M} L_y(\mathcal{D}_k) + L'_y(\mathcal{D}_k) \tag{5.1}$$

$$L_y(\mathcal{D}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(y_{i_k} - \hat{y}_{i_k}\right)^2 \tag{5.2}$$

$$L_y{}'(\mathcal{D}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(y_{i_k} - \hat{y}'_{i_k}\right)^2 \tag{5.3}$$

where N_k is the number of labeled training samples from the k-th source domain; $\hat{y}_{i_k}$ and $\hat{y}'_{i_k}$ are predicted yields by the pair of source-specific predictors for the k-th source domain.

Step 2: In this step, each pair of yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ are trained as discriminators for a fixed feature extractor G_f. By training each pair of yield predictors to increase the discrepancy on target samples, they can detect the target samples outside the support of the source (Figure 5-2. (Step 2)). Specifically, we fix the feature extractor G_f and keep the M pairs of source-specific yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ trainable. The unlabeled target data $\mathcal{D}_t$ are first fed into the feature extractor G_f and then forwarded to all predictors. The yield predictor discrepancy $L_d(\mathcal{D}_t)$ is calculated in L^2 norm between the predicted yield $\hat{y}_{i_k}$ and $\hat{y}'_{i_k}$ for each pair of source-specific predictors G_{p_k} and G'_{p_k} (Eq. (5.5)-(5.6)). Given a target sample out of the support of the source, a large predictor discrepancy is expected (Figure 5-2. (Step 2)). The yield predictors are trained to maximize the predictor discrepancy $L_d(\mathcal{D}_t)$ by minimizing the negative $L_d(\mathcal{D}_t)$ so they can discriminate target samples excluded from the support of the source (Eq. (5.4)). Meanwhile, like Step 1, labeled source data from each source domain are fed into the

network, and the yield prediction loss $L(\mathcal{D}_s)$ on the source samples are calculated (Eq. (5.1)). The source-specific predictors are also trained to minimize the yield prediction loss $L(\mathcal{D}_s)$ to maintain the support of each source. The overall training objective is given as follows (Eq. (5.4)):

$$\min_{\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}} L(\mathcal{D}_s) - L_d(\mathcal{D}_t) \tag{5.4}$$

$$L_d(\mathcal{D}_t) = \sum_{k=1}^{M} L_{d_k}(\mathcal{D}_t) \tag{5.5}$$

$$L_{d_k}(\mathcal{D}_t) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\hat{y}_{i_k} - \hat{y}_{i_k}'\right)^2 \tag{5.6}$$

where N_t denotes the number of unlabeled training samples from the target domain.

Step 3: In this step, the feature extractor G_f is trained to minimize the predictor discrepancy on target samples for fixed yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ (Figure 5-2. (Step 3)). Specifically, the feature extractor G_f is kept trainable while the M pairs of source-specific yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ are fixed. Unlabeled target data $\mathbf{x_t}$ is fed into the network and used to calculate the predictor discrepancy $L_d(\mathcal{D}_t)$ (Eq. (5.5)). During backpropagation, the feature extractor G_f is updated towards minimizing the predictor discrepancy loss (Eq. (5.7)) while the yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ are fixed.

$$\min_{G_f} L_d(\mathcal{D}_t) \tag{5.7}$$

The MMPD model is trained recursively by these three steps until convergence. Overall, the feature extractor G_f and the yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ are trained in an adversarial manner under the condition that the source samples can be predicted correctly. Finally, the

feature extractor G_f is expected to extract task-informative (i.e., informative to the crop yield) and domain-invariant (i.e., small domain shift between source and target domains) features. In other words, source and target distributions are aligned in a task-specific way.

5.2.2 Ensemble Schema

In the testing phase, target samples $\mathbf{x}_i$ are forwarded through the feature extractor G_f and M pairs of source-specific yield predictors $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$. M pairs of predicted yields $\{\hat{y}_{i_k}, \hat{y}'_k\}_{k=1}^{M}$ are given by yield predictors. An ensemble schema was first tested to directly average the *M* pairs of outputs as the final yield prediction but generated unsatisfying results. It was found that different sources may have different relationships with the target, e.g., one source domain might better align with the target domain (Zhao et al., 2020). Therefore, uniformly averaging predictions from all source predictors could result in unsatisfying accuracy.

To address this issue, a weighted ensemble schema was proposed to combine the predictions from each source predictor. In fact, if the target domain and the k-th source domain are well aligned, the k-th pair of source-specific yield predictors will have a low prediction error for the k-th source dataset $\mathcal{D}_k$ and a small predictor discrepancy for the target dataset $\mathcal{D}_t$. Therefore, a weighting schema can be designed based on the prediction errors. Let α_k be the weighting parameter for the k-th source domain with the source-specific prediction error $error_k$ and dis_k be the predictor discrepancy for the target dataset $\mathcal{D}_t$. The weighting parameter α_k is defined as follows:

$$\alpha_k = \frac{\beta}{error_k \, dis_k} \tag{5.8}$$

$$error_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(y_{i_k} - \frac{1}{2}(\hat{y}_{i_k} + \hat{y}'_{i_k})\right)^2 \quad (5.9)$$

$$dis_k = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(\hat{y}_{i_k} - \hat{y}'_{i_k}\right)^2 \quad (5.10)$$

where the source-specific prediction error $error_k$ is defined as the mean squared error (MSE) of the k-th pair of predictors for the k-th source dataset $\mathcal{D}_k$ of size N_k (Eq. (5.9)); the predictor discrepancy dis_k is defined as the mean predictor discrepancy for the target dataset $\mathcal{D}_t$ of size N_t (Eq. (5.10)); β is a tunable parameter to adjust the magnitude of α_k.

Finally, we applied the Softmax function to normalize the weighting parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ (Eq. (5.11)). A weighting vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)$ can be derived and the final prediction $\hat{y}_i$ for the target samples $\mathbf{x}_i$ is a weighted average ensemble of predicted yields from M sources (Eq. (5.12)):

$$w_k = \frac{e^{\alpha_k}}{\sum_{k=1}^{M} e^{\alpha_k}} \quad (5.11)$$

$$\hat{y}_i = \sum_{k=1}^{M}\frac{1}{2}w_k(\hat{y}_{i_k} + \hat{y}'_{i_k}) \quad (5.12)$$

5.2.3 Model Architecture

The architecture of the MMPD model was finalized to have a depth of six after a thorough experimental analysis of accuracy on an independent validation set (Figure 5-1). Specifically, the weight-shared feature extractor G_f consists of three fully connected layers. Each pair of source-specific yield predictor G_{p_k} and G'_{p_k} has two fully connected layers and one output layer. The batch normalization layer (BatchNorm) was used between each hidden layer to address internal covariate shifts and overfitting (Ioffe and Szegedy, 2015). The Rectified Linear Unit (ReLU) was used in each neuron as the activation function. The Adam optimizer was used

to update the model during training. The detailed training process of the proposed MMPD model is illustrated in Table 5.1.

Table 5-1. The Modeling Process of the Proposed MMPD Model

Algorithm 1 Modeling Process of the Proposed Method

procedure DEFINITIONS

$\mathcal{D}_s = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_M\}$: M labeled source domains

$\mathcal{D}_t$: the unlabeled target domain

G_f: the weight-shared feature extractor

$\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$: pairs of source-specific yield predictors for each source domain

Epochs: the total number of training epochs

end procedure

procedure TRAINING PROCESS

1. Initialize G_f with random weights
2. Initialize each pair of source-specific predictors $\{G_{p_k}, G'_{p_k}\}$ with different weights

Repeat steps 3-11 until reaching the Epochs

3. Input $\mathcal{D}_s$ to the MMPD model
4. Calculate the total yield prediction loss $L(\mathcal{D}_s)$ using Eq. (5.1)
5. Update G_f and $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ to minimize $L(\mathcal{D}_s)$
6. Input $\mathcal{D}_s$ and $\mathcal{D}_t$ through the MMPD model
7. Calculate the total yield prediction loss $L(\mathcal{D}_s)$ using Eq. (1); calculate the predictor discrepancy loss $L_d(\mathcal{D}_t)$ using Eq. (5.5)
8. Freeze G_f and update $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ to minimize $L(\mathcal{D}_s)$ and maximize $L_d(\mathcal{D}_t)$
9. Input $\mathcal{D}_t$ through the MMPD model
10. Calculate the predictor discrepancy loss $L_d(\mathcal{D}_t)$ using Eq. (5.5)
11. Freeze $\{G_{p_k}, G'_{p_k}\}_{k=1}^{M}$ and update G_f to minimize $L_d(\mathcal{D}_t)$

end procedure

procedure TESTING PROCESS

1. Input D_s and D_t to the trained MMPD model
2. Calculate the source-specific prediction error $error_k$ on the source dataset D_s for the k-th source using Eq. (5.9)
3. Calculate the predictor discrepancy dis_k on the target dataset D_t for the k-th source using Eq. (5.10)
4. Calculate the weighting parameter α_k of the k-th source using Eq. (5.8)
5. Normalize the weighting parameters using Eq. (5.11)
6. Calculate the final prediction $\hat{y}_i$ for a given target sample $\mathbf{x}_i$ as the weighted average ensemble using Eq. (5.12)

end procedure

5.3 Experimental Setup

We selected the U.S. corn belt and Argentina as the study areas due to the availability of sufficient yield records for model development and validation. Historical yield records in the U.S. corn belt were collected at the county level in 2006-2019 from the USDA National Agricultural Statistics Service (NASS) Database (USDA, 2020b). Similarly, county-level yield records in Argentina from 2006 to 2019 were collected from the Argentina Ministry of Agriculture (Argentine Undersecretary of Agriculture, 2020).

To evaluate the spatial transferability of the proposed MMPD model under different scenarios, three transfer experiments have been designed. In the first experiment, we separated counties in the U.S. corn belt into three domains according to the state-level average yield over the recent ten years (Figure 5-3). As illustrated in Table 5.2, counties in states with an average yield of less than 10.00 t/ha were grouped as the low-yield domain; counties in states with an average yield between 10.00 t/ha and 11.00 t/ha were grouped as the mid-yield domain; counties in states with an average yield higher than 11.00 t/ha were grouped as the high-yield domain. During UDA, three domains would be used as source domains and every single state would be alternatively treated as the target domain. For example, when the target domain is Kansas, labeled training samples in Kansas are first removed from the training set. Then, unlabeled data samples in Kansas and the remaining labeled samples from three source domains are used for model training. Finally, the trained models are evaluated in Kansas in each testing year.

Table 5-2. Multiple domains based on state-level mean yield.

Low-yield states	Mean yield	Mid-yield states	Mean yield	High-yield states	Mean Yield
Kansas (KS)	8.65 t/ha	Michigan (MI)	10.02 t/ha	Minnesota (MN)	11.48 t/ha
North Dakota (ND)	8.78 t/ha	Wisconsin (WI)	10.38 t/ha	Nebraska (NE)	11.50 t/ha
South Dakota (SD)	9.39 t/ha	Ohio (OH)	10.72 t/ha	Illinois (IL)	11.73 t/ha
Missouri (MO)	8.45 t/ha	Indiana (IN)	10.86 t/ha	Iowa (IA)	11.99 t/ha

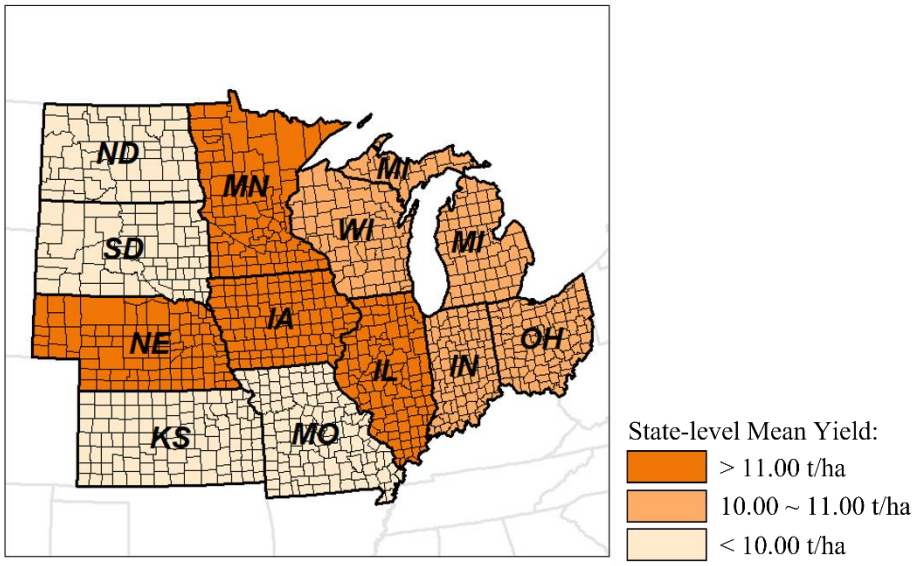

Figure 5-3. Three domains based on the state-level mean yield.

In the second experiment, counties in the U.S. corn belt were grouped into four domains according to eco-regions partitioned by National Ecological Observatory Network (NEON). NEON is a continental-scale research platform for understanding ecosystems (Kampe, 2010). NEON partitions the U.S. into a total of twenty eco-regions, each of which represents different regions of vegetation, landforms, climate, and ecosystem performance. Counties in the U.S. corn belt are in seven NEON eco-regions, including Great Lakes, Prairie Peninsula, Cumberland Plateau, Ozark Complex, Northern Plains, Central Plains, and Southern Plains. Since some eco-regions consist of a very small number of counties, we thus merged small eco-regions and finally resulted in four eco-domains (Figure 5-4). In transfer experiments, each eco-domain would be alternatively treated as the target domain and the other three eco-domains would be treated as sources.

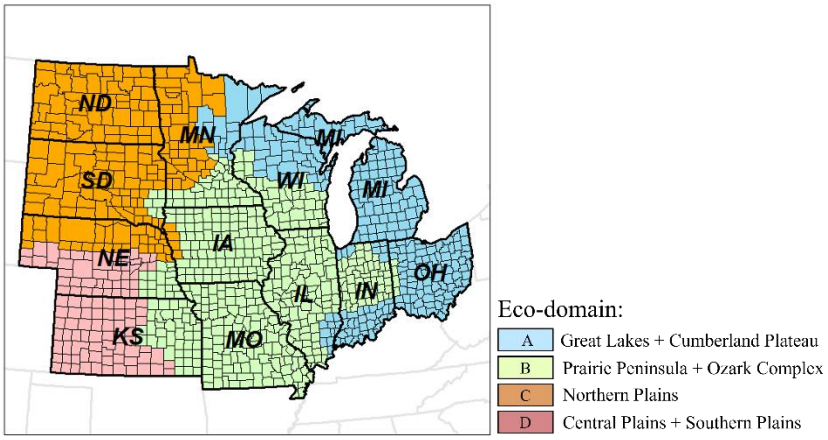

Figure 5-4. Multiple eco-domains partitioned based on NEON eco-regions.

In the third experimental setting, we further evaluate the proposed model through transfer experiments from the U.S. corn belt to Argentina (Figure 5-5). Large domain shifts exist between these two agricultural regions since they are in different hemispheres. The U.S. corn belt has a continental climate while corn-producing areas in Argentina mostly have a humid subtropical climate (Rubel and Kottek, 2011). Following the first experiment, counties in the U.S. corn belt were divided into three domains based on the state-level average yield (Figure 5-3). Labeled samples from these three U.S. source domains and unlabeled samples from Argentina were used for model training. The trained models were evaluated in Argentina in the testing year 2016-2019.

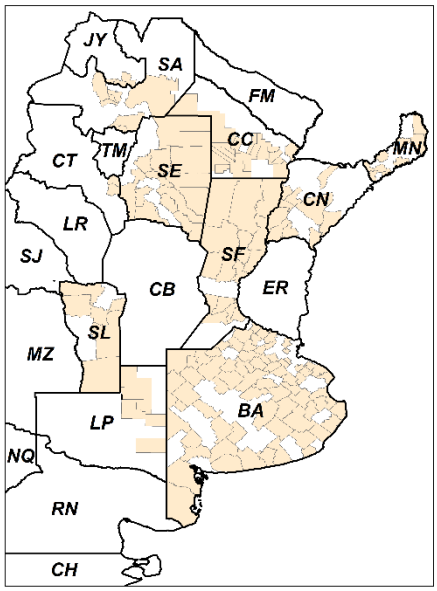

Figure 5-5. Counties with corn yield records in Argentina.

Time-series RS imagery and weather observations were collected as input predictors (Table 5.3). A detailed description of each data source is given in Chapter 2. The MMPD model was compared to three other approaches, including DNN, DANN, and M^{3}SDA. DNN was used as the baseline model and was trained with only labeled source samples. After training, DNN was directly evaluated in the target domain without UDA. When training DANN, multiple sources were grouped into one source domain since DANN is a single-source UDA method. M^{3}SDA was trained following the multi-source setting. In addition, we also evaluated the performance of the Single-source Maximum Predictor Discrepancy (SMPD) model, which is the single-source counterpart of MMPD. SMPD was designed to have a feature extractor and a pair of yield predictors. The final prediction is the average value of the predicted yields from the pair of yield predictors. Like DANN, multiple source domains were grouped into a single source domain when training the SMPD model. We used all preceding years since 2006 for model training and tested models in four testing years 2016–2019. Each model was evaluated based on RMSE and MARE.

Table 5-3. Summary of study areas and data used for model development.

Study Area	Landcover Layer	Predictor Variable	Experiment
The U.S. corn belt	USDA-NASS Cropland Data Layer (CDL)	• EVI, NDWI, GCI from MODIS MCD43A4 • LSTday, LSTnight from MODIS MYD11A2 • Tmean, Tmax, PPT from PRISM	• Experiment 1: **Source**: Three source domains based on state-level average yield in the U.S. corn belt → **Target**: One state in the U.S. corn belt. • Experiment 2: **Source**: Three NEON eco-regions in the U.S. corn belt → **Target**: One NEON eco-region in the U.S. corn belt. • Experiment 3: **Source**: Three source domains based on state-level average yield in the U.S. corn belt → **Target**: Argentina.
Argentina	MODIS Land Cover Type product (MCD12Q1 v6)	• EVI, NDWI, GCI from MODIS MCD43A4 • LSTday, LSTnight from MODIS MYD11A2 • Tmean, Tmax from ERA5 • PPT from CHIRPS	

5.4 Results and Discussion

5.4.1 Transfer experiments among U.S. States

We first evaluated the MMPD model based on three source domains in the U.S. corn belt which are divided based on the state-level average yield (Figure 5-3). The evaluation results averaged over 2016-2019 in each target state are reported in Table 5.4. The best performer in each case is highlighted in bold. MMPD was observed to outperform other models in each target state. Specifically, DNN performed well in some states but had low accuracy in others, especially in low-yield states such as Kansas and North Dakota. The reason is that most training samples are from the mid-yield and high-yield states which would introduce bias in training DNN and limit its generalizability to low-yield states. Through UDA, DANN and M^3SDA had improved performance in a few target states (i.e., Missouri and Ohio) but performed worse than DNN in most cases. These results indicate that merely aligning feature distributions in source

and target domains without considering the yield response in the target domain might invalidate the predictor training since the samples in the target domain might be mistakenly aligned to target samples with different yield levels. The SMPD model had comparable corn yield prediction in several states (i.e., Ohio and Wisconsin) but had poor performance in most cases. For example, SMPD performed poorly in high-yield states, such as Illinois and Iowa. This demonstrated that grouping all labeled samples into one source could increase the difficulty for the model to learn from data samples collected from highly heterogenous regions. Instead, MMPD outperformed SMPD and other models with better prediction in all cases. With MPD and multi-source strategy, MMPD could align target samples to the most relevant source and extract informative features for the yield prediction task.

Table 5-4. Average evaluation results of RMSE (t/ha) and MARE in each target state in testing years 2016-2019.

Target	#Counties	DNN		DANN		M^{3}SDA		SMPD		MMPD	
		RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE
IL	98	1.30	8.77%	1.60	10.92%	1.66	11.31%	1.83	12.42%	**1.25**	**8.40%**
IN	85	0.87	6.07%	0.91	6.54%	1.05	7.42%	1.08	8.02%	**0.84**	**5.96%**
IA	99	1.26	8.46%	1.88	13.30%	1.99	13.57%	1.96	13.81%	**1.20**	**8.01%**
KS	71	1.41	14.51%	1.59	16.30%	1.74	19.38%	1.70	17.95%	**1.26**	**12.87%**
MI	59	1.14	10.34%	1.09	9.80%	1.22	10.66%	1.11	9.88%	**1.06**	**9.63%**
MN	76	1.53	10.75%	1.63	11.43%	1.27	8.89%	1.67	12.23%	**1.09**	**7.71%**
MO	70	1.51	13.23%	1.16	9.83%	1.23	10.98%	1.47	13.12%	**1.11**	**9.70%**
NE	81	1.40	9.16%	1.48	10.34%	1.42	9.67%	1.98	14.11%	**1.35**	**8.48%**
ND	43	1.20	12.94%	1.37	15.19%	2.05	25.92%	1.22	13.78%	**1.11**	**12.16%**
OH	77	1.14	8.84%	1.08	8.61%	1.13	8.35%	1.12	8.48%	**0.92**	**7.38%**
SD	47	1.49	12.41%	1.27	11.03%	1.39	14.14%	1.22	9.84%	**1.10**	**9.04%**
WI	57	1.22	9.29%	1.08	8.14%	1.34	10.08%	1.04	7.81%	**0.94**	**7.20%**

To evaluate whether the methods were statistically different on the reported MARE, a paired sample t-test was used (Mishra et al., 2019). The statistical tests between the evaluation results of MMPD and each comparison model were performed. A t-test is a statistical test that compares the means of two samples. In this experiment, we compared the means of MARE of each model in all testing years. Since each experiment was repeated five times in each state,

there were totally 60 pairs of samples in each t-test. As shown in Table 5-5, the accuracy improvement obtained by the proposed MMPD was statistically significant.

Table 5-5. The paired sample t-test of the transfer experiments among U.S. States between the MARE of each comparison model and MMPD.

Model	t	*p*-value
MMPD vs DNN	17.379	0.000
MMPD vs. DANN	21.640	0.000
MMPD vs. MSDA	23.340	0.000
MMPD vs SMPD	28.856	0.000

Furthermore, we showed the density scatter plots for each model in four representative target states to compare the agreement between the reported and the predicted yields in the average of four testing years (Figure 5-6). The best agreement was again observed from the MMPD model. It is also observed that all models show top performance in Indiana. Specifically, the non-UDA model DNN illustrated the best prediction results in Indiana. This indicates that Indiana has the smallest domain shift, which has been further evidenced by the best performance among all states for almost all UAD models (except M^3SDA). However, DNN performed poorly in South Dakota with obvious underestimations (Figure 5-6 (a3)). The reason is that South Dakota is located on the boundary of the corn belt and has a large domain shift with respect to other states. DANN and M^3SDA relatively underperformed in predicting the yields in selected states with larger disagreement with the reported yields (Figure 5-6 (b)&(c)) since they merely matched feature distributions in source and target domains without considering target yield response. SMPD had better agreement than DANN and M^3SDA in South Dakota and Wisconsin but had no obvious improvement in comparison with DNN (Figure 5-6). The MMPD model outperformed other models and had a better agreement in chosen target domains (Figure 5-6 (e)). Moreover, we noticed that all UDA models tended to underestimate corn yield in Iowa. Iowa is

the top corn-producing state in the U.S. with lots of high-yield counties. The UDA models' prediction results indicate that the prior information or local training samples from the target domain may be still needed when large biases exist between different domains. Even though, as shown in Figure 5-6, MMPD is still the best-performing model in Iowa with the least underestimation.

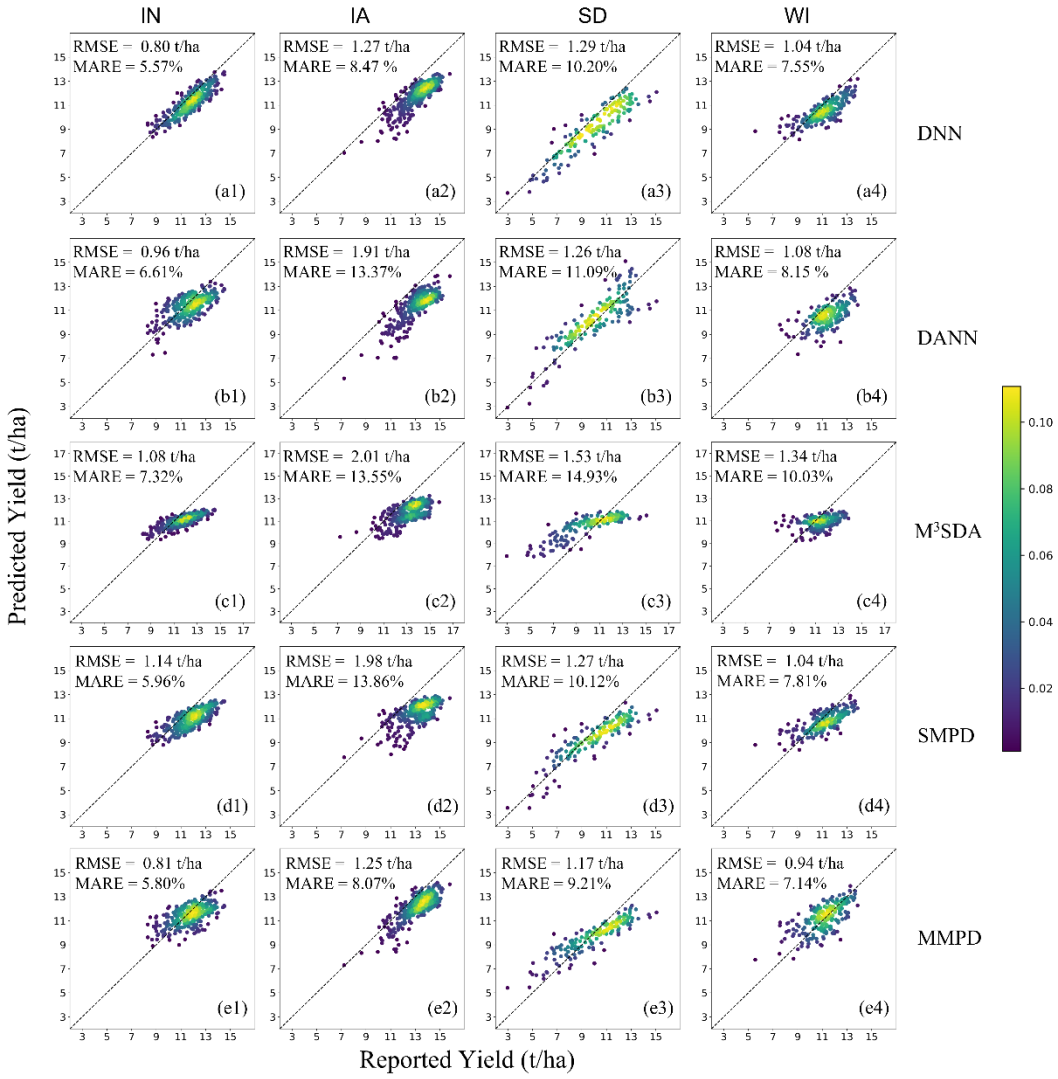

Figure 5-6. The density scatter plots of reported yields vs. predicted yields in 2016-2019 of (a) DNN, (b) DANN, (c) M^{3}SDA, (d) SMPD, (e) MMPD in (1) IN, (2) IA, (3) SD, (4) WI.

Finally, we presented the absolute error maps averaged over four testing years for each model, in which darker color represents a larger error. Still, the proposed MMPD model was

observed to have better spatial transferability than the other UDA models. In concrete, M^3SDA and DANN were observed to have spatial clustering of large errors in the east of IA (Figure 5-7 (b2)&(c2)) and central South Dakota (Figure 5-7 (c3)). Meanwhile, in Indiana (Figure 5-7 (b1)&(c1)) and Wisconsin (Figure 5-7 (b4)&(c4)), M^3SDA and DANN tended to make larger errors than the DNN model. SMPD had solved the problem of the spatial cluster of big errors but still had big prediction errors in Iowa and South Dakota (Figure 5-7 (d2)&(d3)). Comparatively, MMPD had better addressed the domain shift issue with smaller prediction errors in all selected target states.

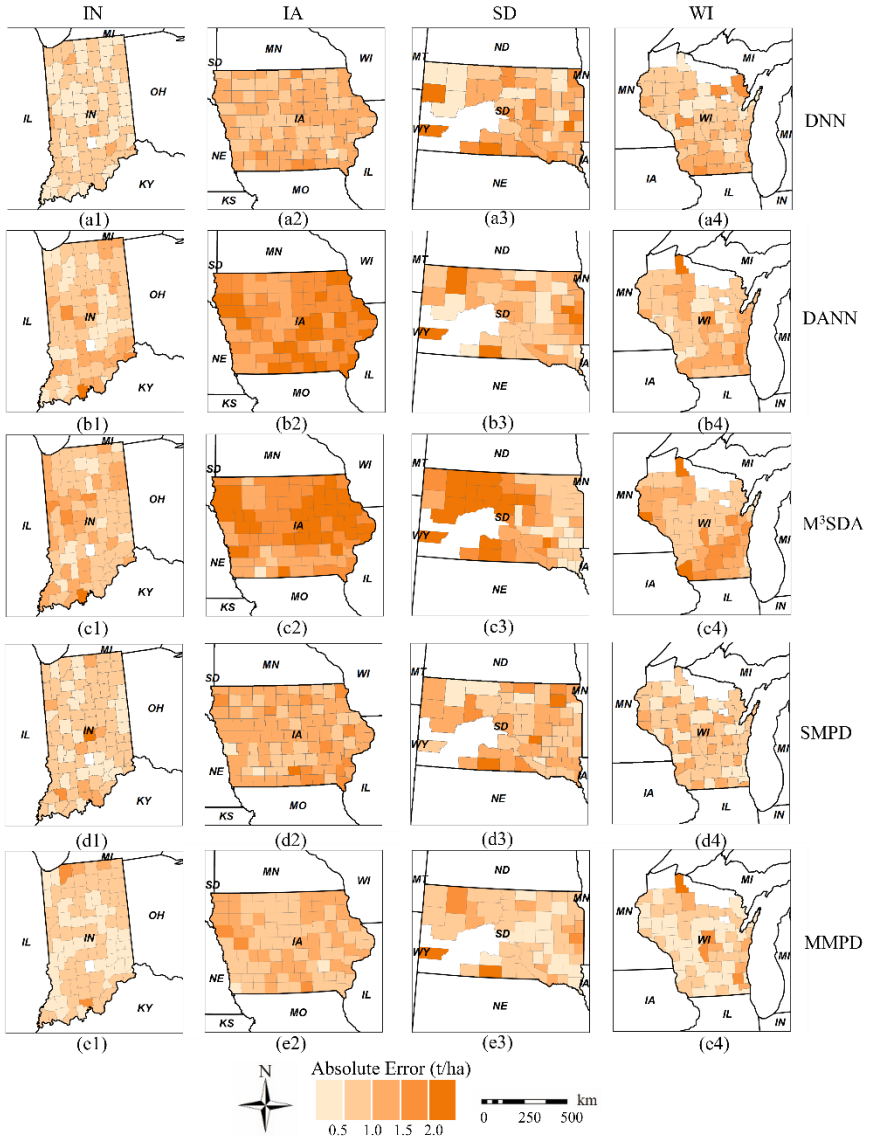

Figure 5-7. The average absolute error maps in 2016-2019 for model (a) DNN, (b) DANN, (c) M^3SDA, (d) SMPD, (e) MMPD in (1) Indiana, (2) Iowa, (3) South Dakota, (4) Wisconsin.

5.4.2 Transfer experiments among U.S. Ecoregions

In the second experiment, we evaluated MMPD via UDA among four eco-domains in the U.S. corn belt (Figure 5-4). The evaluation results of four-year average yield predictions of 2016-2019 are given in Table 5-5 with the best performer highlighted in bold.

Table 5-6. Average evaluation results of RMSE (t/ha) and MARE in each target eco-domain in testing years 2016-2019.

Target	#Counties	DNN	DANN	M^3SDA	SMPD	MMPD

		RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE
A	204	1.15	9.18%	1.04	7.84%	1.25	10.24%	1.10	8.69%	**0.97**	**7.76%**
B	431	1.35	9.76%	**1.30**	**9.11%**	2.19	15.43%	2.16	15.39%	1.90	13.51%
C	145	1.56	13.31%	1.56	14.53%	1.77	15.60%	1.53	13.12%	**1.47**	**13.02%**
D	78	1.36	12.14%	1.32	11.22%	2.12	17.56%	1.32	10.43%	**1.18**	**10.27%**

As shown in Table 5-6, DNN performed poorly in the target eco-domain B, eco-domain C, and eco-domain D with RMSE over 1.30 t/ha due to the existence of domain shift. DANN slightly outperformed DNN in eco-domain A, eco-domain B, and eco-domain D due to the effects of UDA. However, it was observed that M^{3}SDA again failed to effectively address the domain shift issue and performed poorly in all eco-domains. On the other hand, SMPD performed better than DANN in eco-domain C and D but had lower accuracy in eco-domain A and eco-domain B (Table 5-6). MMPD, however, improved the prediction accuracy and outperformed other models in eco-domain A, eco-domain C, and eco-domain D. However, both SMPD and MMPD had worse performance than DNN and DANN in eco-domain B. To evaluate whether the methods were statistically different on the reported MARE, a paired sample t-test was used. In this experiment, we compared the means of MARE of each model in all testing years. Since each experiment was repeated five times in each eco-domain in each year, there were totally 80 pairs of samples in each t-test (Table 5-7). Due to its comparatively poor performance in Ecoregion B, the MMPD didn't perform significantly better than DNN or DANN. However, the results still demonstrated that the accuracy improvement obtained by the proposed MMPD was significantly better than M^{3}SDA and SMPD.

Table 5-7 The paired sample t-test of the transfer experiments among U.S. Ecoregions between the MARE of each comparison model and MMPD.

Model	**t**	***p*-value**
MMPD vs DNN	-0.297	0.767
MMPD vs. DANN	-2.710	0.007
MMPD vs. **M^{3}SDA**	20.007	0.000
MMPD vs SMPD	15.009	0.000

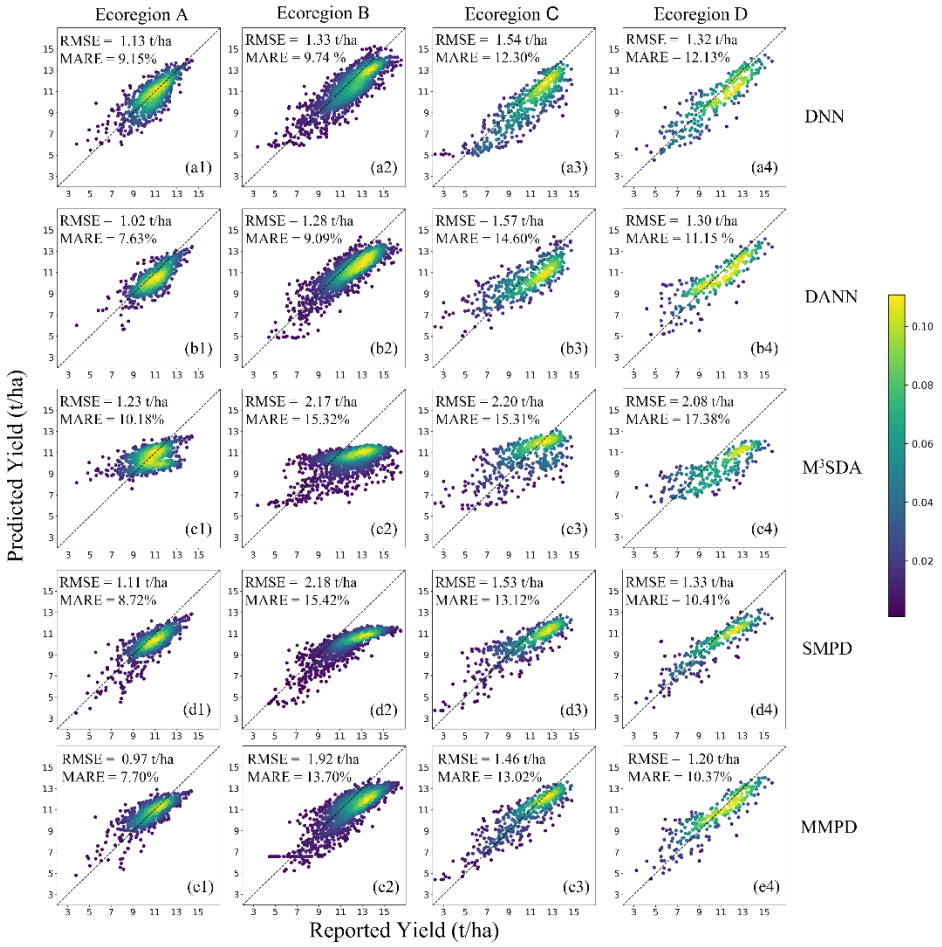

Figure 5-8. The density scatter plots of reported yields vs. predicted yields in 2016-2019 of (a) DNN, (b) DANN, (c) M^{3}SDA, (d) SMPD, (e) MMPD in (1) eco-domain A, (2) eco-domain B, (3) eco-domain C, and (4) eco-domain D.

Similarly, Figure 5-8 illustrated the density scatter plots of reported yields vs. predicted yields in each target eco-domain in the four-year average of 2016-2019 for all prediction models. The proposed MMPD model was observed to have the best agreement in most target eco-domains though its prediction is inferior to DNN and DANN for eco-domain B. It was further observed that the scatteredness of the predictions highly depended on the eco-domains as evidenced by the high compactness of eco-domain A and vs high scatteredness of eco-domain C and eco-domain D. This may indicate the homogeneity differences among different regions and

the need for further subdividing the domains for training and prediction.

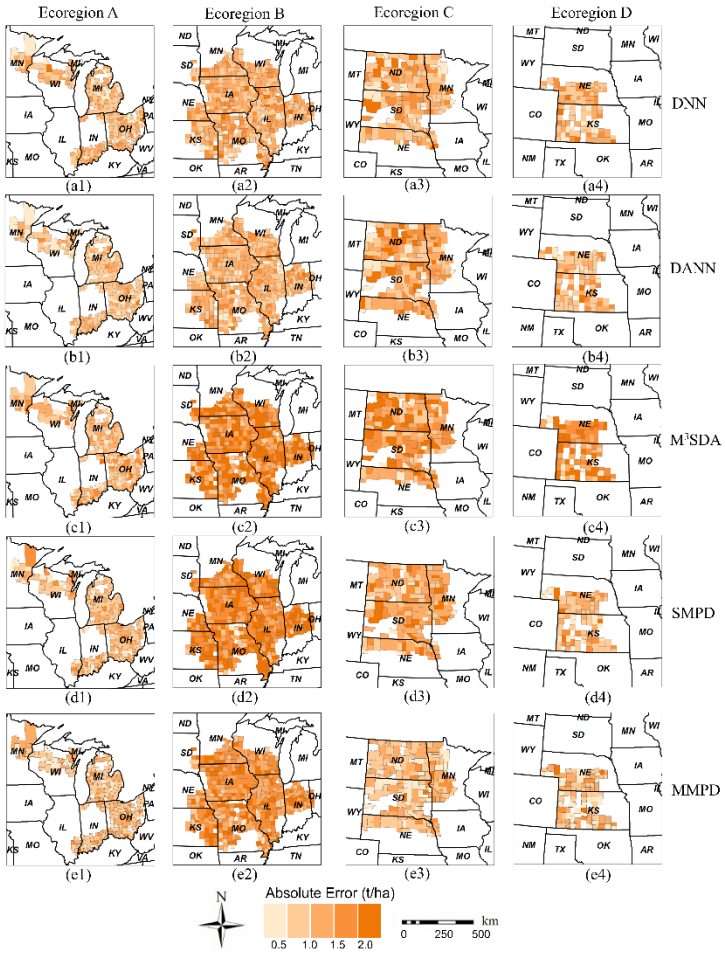

Figure 5-9. The average absolute error maps in 2016-2019 for model (a) DNN, (b) DANN, (c) M^3SDA, (d) SMPD, (e) MMPD in (1) eco-domain A, (2) eco-domain B, (3) eco-domain C, and (4) eco-domain D.

Finally, we presented the four-year average absolute error maps for each model (Figure 5-9). Again, the proposed MMPD model had better spatial transferability than other models in most cases (Figure 5-9 (e)). However, it was observed that the overall prediction errors of MMPD were larger than DNN and DANN in eco-domain B (Figure 5-9 (a2), (b2), (d2), and (e2)), which were consistent with the results shown in Table 5-5. As shown in Figure 5-9, the UDA models M^3SDA, SMPD, and MMPD all showed degraded performance in eco-domain B. It was noticed that the number of counties in eco-domain B significantly exceeds the number of counties in other eco-domains as shown in Table 5-5. This meant smaller training sets of three

these source domains were used as compared to the target domain B. Therefore, when conducting UDA to eco-domain B, each domain-specific yield predictor of MMPD was not sufficiently trained. This indicates that large source domains are required to guarantee the success of the proposed MMPD model. Even though, MMPD still outperformed SMPD by a large margin in eco-domain B since negative interference among different source samples was well handled by MMPD.

5.4.3 Transfer experiments from the U.S. corn belt to Argentina

The model performance was further tested for UDA from the U.S. corn belt to Argentina. A large domain shift exists between the U.S. corn belt and Argentina since they are in different hemispheres with different climates. The U.S. corn belt was chosen as the source since it has more labeled data samples for model training. Specifically, DNN was trained in the U.S. corn belt and evaluated in Argentina without UDA or TL. DANN and SMPD were trained using the whole U.S. corn belt as a single source domain. To train M^{3}SDA and MMPD, counties in the U.S. corn belt were divided into three source domains based on the state-level average yield for multi-source UDA (Figure 5-3). The evaluation results of each model for 2016-2019 are reported in Table 5-6 with the best results highlighted in bold for each year.

Table 5-8. Average evaluation results of RMSE (t/ha) and MARE in domain adaptation from the U.S. corn belt to Argentina in each testing year 2016-2019.

Testing year	#Counties	DNN		DANN		M^{3}SDA		SMPD		MMPD	
		RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE	RMSE	MARE
2016	122	1.87	18.86%	1.64	16.78%	1.92	19.82%	1.79	18.70%	**1.41**	**15.12%**
2017	116	1.86	18.08%	1.79	17.81%	1.52	15.88%	1.66	16.30%	**1.48**	**14.43%**
2018	113	2.01	27.08%	1.88	25.32%	1.71	22.47%	1.96	26.76%	**1.60**	**21.00%**
2019	118	2.46	24.92%	2.26	21.94%	2.02	20.92%	2.21	22.63%	**1.98**	**19.23%**

To evaluate whether the methods were statistically different on the reported MARE, a paired sample t-test was used. The statistical tests between the evaluation results of MMPD and

each comparison model were performed, and the results are shown in Table 5-9. In this experiment, we compared the means of MARE of each model in all testing years. Since each experiment was repeated five times in each year, there were totally 20 pairs of samples in each t-test. The results demonstrated that the prediction accuracy obtained by the proposed MMPD was significantly better than the other comparison models in the transfer experiments from the U.S. corn belt to Argentina.

Table 5-9. The paired sample t-test of the transfer experiments from the U.S. corn belt to Argentina between the MARE of each comparison model and MMPD.

Model	t	*p*-value
MMPD vs DNN	20.210	0.000
MMPD vs. DANN	13.765	0.000
MMPD vs. M^{3}SDA	12.645	0.000
MMPD vs SMPD	15.091	0.000

Due to the big domain shift between the U.S. corn belt and Argentina, DNN performed poorly in all testing years, especially in 2018, when the corn harvest in Argentina was hit by the worst drought in half a century ("Buenos Aires Times | Corn to surpass soybean production this year, says Argentina," 2019). By aligning the feature distributions in source and target domains, all UDA models showed different levels of prediction accuracy improvement as compared to the DNN model but had different stabilities. For example, in 2018, both DANN and M^{3}SDA improved the prediction accuracy and reduced the RMSE by large margins compared to DNN. However, in 2017, DANN had barely improved its accuracy in comparison with DNN. Also, in 2016, M^{3}SDA had worse performance than DNN and made about 1.00% more MARE. SMPD was observed to outperform DNN in all testing years but underperformed DANN and M^{3}SDA in some testing years. In comparison, the proposed MMPD outperformed DNN and all other UDA models in all testing years (Table 5-6).

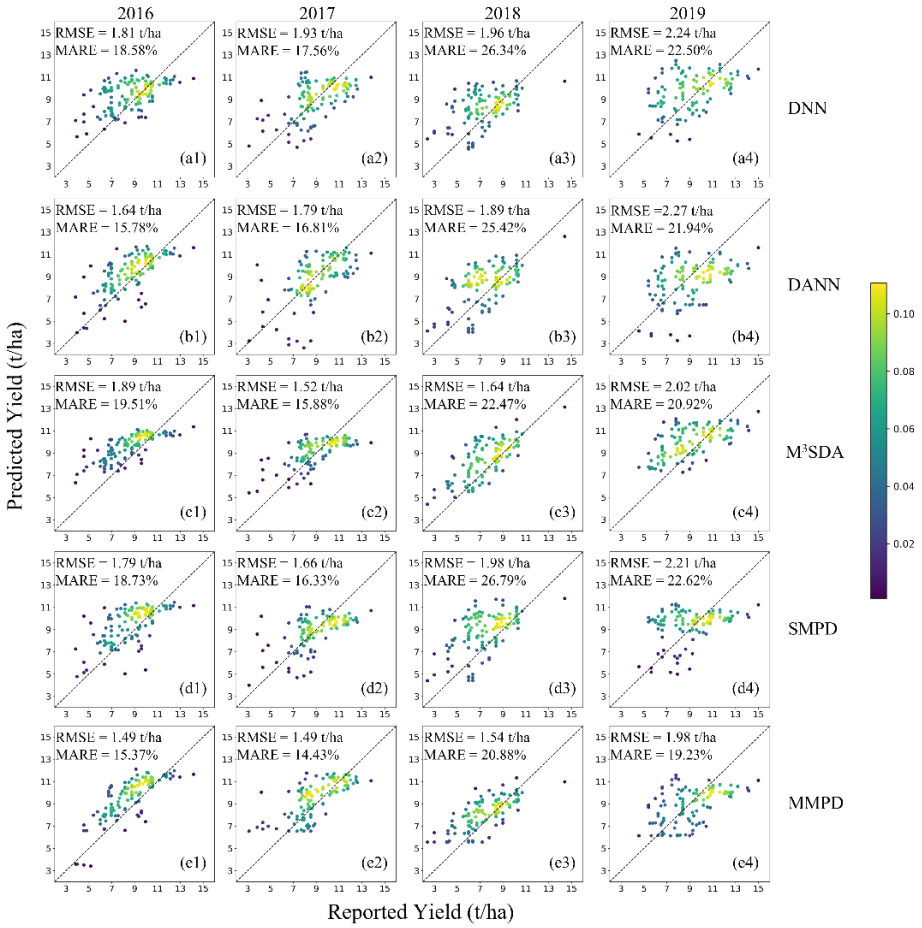

Figure 5-10. The density scatter plots of reported yields vs. predicted yields in Argentina of (a) DNN, (b) DANN, (c) M^{3}SDA, (d) SMPD, (e) MMPD in the testing year (1) 2016, (2) 2017, (3) 2018, (4) 2019.

Fig 5-10 shows the density scatter plots of each model to demonstrate the agreement between the reported and predicted corn yields in each testing year. Again, the predicted yields by DNN were in poor agreement with the reported yields in most testing years due to the domain shift. As shown in Figure 5-10, different models have different levels of prediction bias in different years. In particular, DNN significantly underestimated corn yield in 2017 and 2019 (Figure 5-10 (a2)-(a4)) while corn yield was significantly overestimated by DANN in 2017, M^{3}SDA in 2016, 2018, and 2019, SMPD in 2016, 2018, and 2019. However, MMPD further improved the prediction accuracy and achieved the best agreement in all four testing years with the least estimation bias (Figure 5-10 (e1)-(e4)).

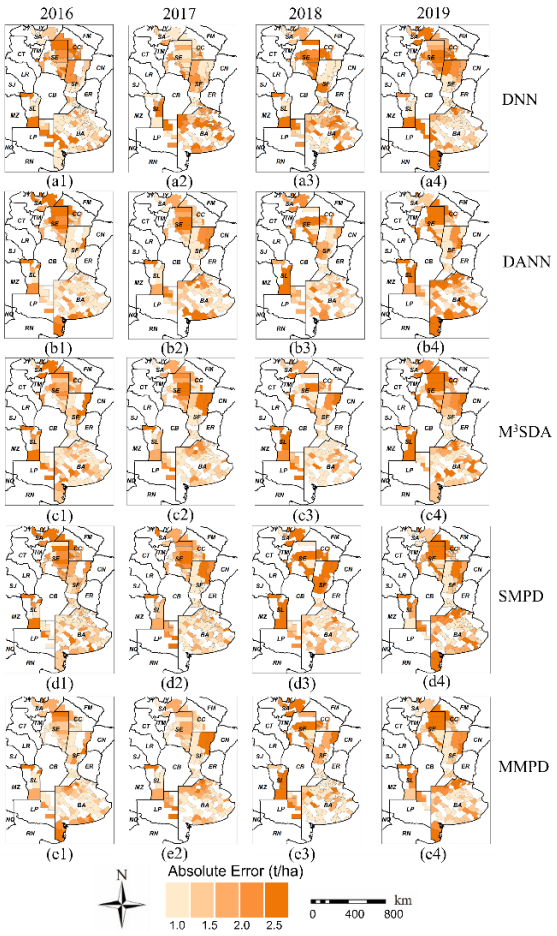

Figure 5-11. The average absolute error maps in Argentina of model (a) DNN, (b) DANN, (c) M^3SDA, (d) SMPD, (e) MMPD in the testing year (1) 2016, (2) 2017, (3) 2018, (4) 2019.

Finally, Figure 5-11 illustrates the corresponding absolute error maps for all models in each testing year. It was observed that DNN constantly made large errors across Argentina, especially in provinces Santiago del Estero (SE), Santa Fe (SF), and Buenos Aires (BA). DANN had reduced errors in SF and BA in 2018 (Figure 5-11 (b3)) but no obvious improvement was observed in other testing years. Similarly, the performance of M^3SDA was not stable. For example, in comparison with DNN, M^3SDA had fewer errors in SF in 2018 (Figure 5-11 (c3)) but had more errors in the same region in 2017 and 2019 (Figure 5-11 (c2)&(c4)). The SMPD model had a decent performance in BA but performed poorly in SE (Figure 5-11 (d)). Again, the proposed MMPD model outperformed the other models with obvious improvements in each

testing year. For example, MMPD constantly reduced errors in BA and improved yield prediction in southern SE and SF (Figure 5-11 (e)). Also, MMPD accurately estimated the corn yields in SF with all absolute errors less than 1.50 t/ha in 2018 (Figure 5-11 (e3)). Furthermore, it was noted that SE and northern SF constantly had large errors. These regions have a very low corn yield with an average yield below 4.00 t/ha. As a result, it was challenging to align these data samples to the source domains since most U.S. counties have a corn yield larger than 6.00 t/ha. Therefore, all UDA models made large prediction errors in these regions.

5.4.4 t-SNE Visualization of Feature Distribution

To provide a visual insight into the effects of UDA by each model, we visualized the feature distributions of the input features as well as the extracted features by each UDA model using the t-distributed Stochastic Embedding (t-SNE) algorithm (Maaten and Hinton, 2008). As shown in Figure 5-12 - 5-14, one example for each UDA scenario in the testing year 2019 was presented since similar results were obtained in other cases or other testing years.

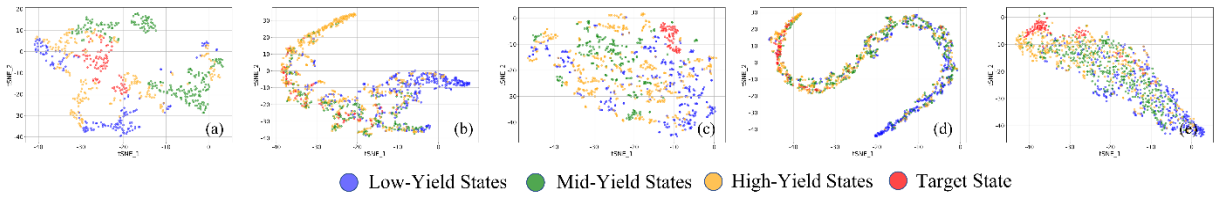

Figure 5-12. The t-SNE visualization of (a) input features, and extracted features from (b) DANN, (c) M^{3}SDA, (d) SMPD, and (e) MMPD in three source domains (i.e., U.S. low-yield domain, mid-yield domain, high-yield domain) and the target domain (i.e., Iowa) in the testing year 2019.

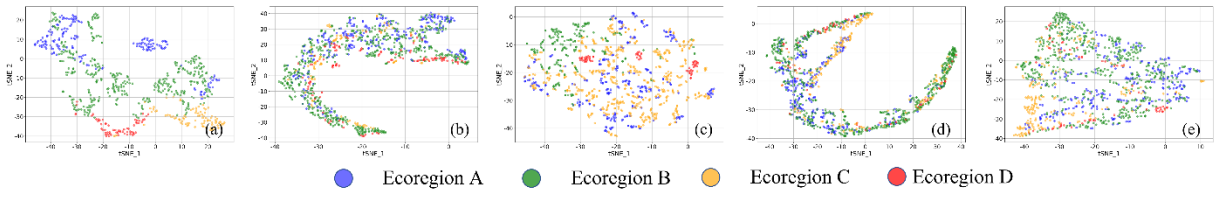

Figure 5-13. The t-SNE visualization of (a) input features, and extracted features from (b) DANN, (c) M^{3}SDA, (d) SMPD, and (e) MMPD in three source domains (i.e., eco-domain A-C) and the target domain (i.e., eco-domain D) in the testing year 2019.

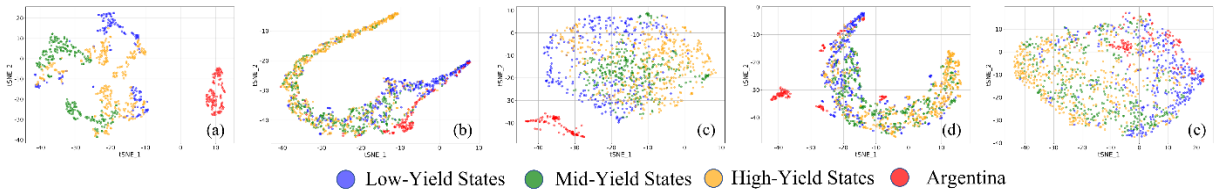

Figure 5-14. The t-SNE visualization of (a) input features, and extracted features from (b) DANN, (c) M^{3}SDA, (d) SMPD, and (e) MMPD in three source domains (i.e., U.S. low-yield domain, mid-yield domain, high-yield domain) and the target domain (i.e., Argentina) in the testing year 2019.

The goal of UDA is to blend features from different domains in a uniform distribution in the feature space. For given input features, different t-SNE visualization results were illustrated in Figure 5-12 – 5-14 for different source and target domains. For example, as shown in Figure 5-14 (a), the t-SNE transformed features of the domains of the U.S. corn belt and Argentina clearly illustrated domain shift between source and target domains as well as among different source domains, as those domain features were separately clustered. After domain adaptation with different UDA methods, the domain shift has been reduced to different degrees as shown in Figure 5-12 – 5-14 (b)-(e).

Specifically, DANN could well merge the source and target samples. However, since DANN addresses the domain shift by directly matching feature distributions in source and target domains, the wrong alignment would happen. For example, when Iowa was the target domain, quite a few target samples were aligned to the mid-yield domain by DANN although Iowa is a high-yield state (Figure 5-12 (b)). Similarly, M^{3}SDA tries to reduce the domain discrepancy by matching moments across all pairs of source and target domains. Therefore, it was observed that feature distributions in the target domain were aligned closer to those in source domains by M^{3}SDA, but individual target samples were not uniformly mixed with source samples. For example, when eco-domain D was the target domain, M^{3}SDA aligned target samples closer to the source samples but failed to dismantle the cluster of target samples (Figure 5-13 (c)). This

explains why M^{3}SDA performed poorly in most cases. SMPD better aligned the target domain to source domains while there were still some extracted target feature samples located far from the support of source domains (Figure 5-14 (d)). For example, when Argentina was the target domain, a cluster of target samples extracted by SMPD was observed to be outside the support of either of the source domain. Moreover, for single-source UDA methods DANN and SMPD, it was noted that they tended to merge the source and target samples to a narrow space (Figure 5-12 – 5-14 (b)&(d)). It indicated that samples from different sources have also been tightly aligned by DANN and SMPD, which could cause negative interference. On the other hand, the t-SNE transformed features of MMPD have shown the best adaptation patterns (Figure 5-12 – 5-14 (e)). Through maximum predictor discrepancy and multi-source UDA, MMPD aligned target samples to the most similar source domain by extracting domain-invariant and task-informative features. For example, as shown in Figure 5-12 (e), target samples from Iowa have been mostly aligned with the high-yield domain since counties in Iowa mostly have high corn yields. Similarly, when Argentina was the target domain, most target samples have been matched to the low-yield and mid-yield domains (Figure 5-14 (e)) in the U.S. since the corn yields in Argentina are mostly located in the range.

5.5 Summary

In this study, we proposed a multi-source UDA method named MMPD for county-level corn yield prediction based on time-series RS imagery and weather variables (Ma et al., 2023; Ma and Zhang, 2021). The proposed MMPD model aims to reduce the domain shift between source and target domains and accurately predict corn yield in the target domain without using labeled data from the target domain. By using the MPD, MMPD was trained to align source and target domains by considering crop yield response in the target domain based on task-specific

regression models. Also, the multi-source UDA strategy was adopted in MMPD to avoid negative interference among source samples from heterogeneous regions. Experiments on three UDA scenarios in the U.S. corn belt and Argentina have been conducted to evaluate the model performance. It was observed that MMPD outperformed representative single-source and multi-source UDA methods. Also, the t-SNE visualization analysis showed that MMPD not only reduced the domain shift between source and target domains but also matched the target samples to the most similar source domain.

The main contributions of this work are summarized as follows:

- A multi-source UDA method, i.e., MMPD, has been developed for corn yield prediction based on RS imagery and weather variables.
- Instead of aligning feature distributions only, the MPD was used to align source and target domains by considering crop yield response in the target domain.
- The strategy of multi-source UDA was adopted for the first time in the yield prediction tasks to avoid negative interference among labeled data samples from heterogeneous spatial regions and better leverage data from multiple source domains.
- The proposed MMPD model outperformed commonly used supervised learning, single-source UDA, and multi-source UDA models with improved spatial transferability in different scenarios across the U.S. corn belt and Argentina for multiple testing years.

CHAPTER 6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

Recently, ML and DL models have been explored for corn yield prediction. Despite the success, there are still two major limitations of existing DL-based crop yield prediction models. First, existing models mainly focus on predicting the crop yield without providing any information about the predictive uncertainty which is important for quantifying the confidence interval of the prediction. Second, data-driven ML and DL models require a large amount of data samples with labels (i.e., yield records) for model training and tend to have low spatial transferability due to domain shifts among different regions. In this dissertation, we tried to address these two major limitations through Bayesian inference and UDA. Specifically, the contributions of this dissertation are summarized as follows:

- In Chapter 3, we tried to address the first challenge by introducing Bayesian inference into DL and proposed a BNN model for corn yield prediction and uncertainty estimation. The proposed BNN model could accurately predict county-level corn yield and outperforms other ML and DL models. Also, accurate corn yield prediction could be made by the proposed BNN model in August, which is about two months ahead of the harvest. Moreover, it has been proven that predictive uncertainty has a strong correlation with prediction error. It means that the predictive uncertainty can be used to quantify the quality of the prediction even without the ground truth yield records. Finally, the potential sources of predictive uncertainty have been analyzed. It has been found that observation noises and environmental stresses, such as heat stress

and water stress, could potentially increase the uncertainty in corn yield prediction (Ma et al., 2021a).

- In Chapter 4, we tried to address the second challenge by using the UDA strategy. Two adversarial domain adaptation models were proposed for county-level corn yield prediction based on RS images and weather variables. The proposed ADANN and BDANN have been proven to have better spatial transferability across spatial regions and outperformed other supervised learning models and the original DANN in transfer experiments. The t-SNE visualization showed that ADANN and BDANN were able to effectively reduce the domain shift and well align the source domain and the target domain (Ma et al., 2021b; Ma and Zhang, 2022).
- In Chapter 5, we further addressed the existing issue with current single-source UDA methods and proposed a multi-source UDA method, i.e., MMPD, for corn yield prediction based on RS and weather variables. Instead of aligning feature distributions only, the MPD was used to align source and target domains by considering crop yield response in the target domain. Also, the strategy of multi-source UDA was adopted for the first time in the yield prediction tasks to avoid negative interference among labeled data samples from heterogeneous spatial regions and better leverage data from multiple source domains. The proposed MMPD model outperformed commonly used supervised learning, single-source UDA, and multi-source UDA models with improved spatial transferability in different UDA scenarios across the U.S. corn belt and Argentina for multiple testing years.

6.2 Future Work

The future work can be focused on the following research directions:

- **Multiple Instance Regression**: In this dissertation, the RS images and weather variables were preprocessed by first masking out irrelevant pixels based on a cropland layer and then aggregating RS images and weather variables to the county level by calculating their mean values in each county. However, directly aggregating all pixels within each county would cause information loss. To fully utilize the detailed information in RS images, multiple instance regression (MIR) (Wagstaff et al., 2008) is a promising strategy. In the scenario of county-level crop yield prediction, instead of purely using the mean value of image pixels to represent the county, MIR considers each county as a bag, its crop yield as the bag label, and pixel-level observations within the county as instances. As a result, RS images are organized as bags of instances, with a single label (i.e., crop yield records) applied to each bag. With multiple instances in each bag, more detailed information has been kept and enabled the regressor to better associate the RS images with the crop yield.
- **Partial Domain Adaptation**: Our current studies on UDA generally assume identical label space across different domains. This assumption can be invalid in scenarios such as crop yield prediction since the yield levels can be very different from region to region. As a result, a negative transfer could happen, and target samples can be mistakenly aligned to source samples with very different yields. To alleviates negative transfer caused by the mismatch of label spaces, partial domain adaptation (PDA) can be a promising strategy (Cao et al., 2018). Instead of matching the whole source domain and the target domain, PDA tries to down-weight the outlier source samples during model training. Therefore, source samples that are within the label space of the target samples will be given large weights during training while potential

outlier source samples will be given small weights. As a result, the negative transfer can be effectively alleviated while the positive transfer can be prompted.

- **Knowledge-guided Machine Learning**: Fixed error patterns have been observed in certain areas of our research activities. For example, in both the supervised learning experiments and transfer experiments, models tend to underestimate the corn yield in Iowa. The reason is that Iowa is the top corn-producing state in the U.S. Since the training dataset is not uniformly distributed and has comparatively fewer high-yield samples, either supervised learning models or UDA models would biasedly underestimate the high-yield counties such as those from Iowa. To address this issue and avoid biased corn yield prediction, a promising way is to use knowledge-guided ML by introducing prior knowledge to ML. With prior knowledge such as the historical yield level in certain areas, the ML models can reduce potential biases in the prediction.
- **Explainable and Interpretable Transfer Learning:** The explainability and interpretability of TL and UDA models is essential for understanding how they work and for identifying the factors that influence their transferability and performance. Also, with interpretable models, the input feature variables with high transferability can be identified, which will provide guidance for feature selection and model architectures. However, most of existing TL and UDA models in agriculture lack interpretability. Existing studies of TL in agriculture focus on training strategies and model performance. The lack of interpretability of TL models in agriculture can limit their usefulness and adaptation. Therefore, improving the interpretability of TL

models in agriculture is crucial, particularly in the context of climate change, to enable better utilization of these models in tackling unforeseen issues.

REFERENCE

Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W., He, Y., 2019. Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. Comput. Electron. Agric. 167, 105091. https://doi.org/10.1016/j.compag.2019.105091

Archontoulis, S. V., Castellano, M.J., Licht, M.A., Nichols, V., Baum, M., Huber, I., Martinez-Feria, R., Puntel, L., Ordóñez, R.A., Iqbal, J., Wright, E.E., Dietzel, R.N., Helmers, M., Vanloocke, A., Liebman, M., Hatfield, J.L., Herzmann, D., Córdova, S.C., Edmonds, P., Togliatti, K., Kessler, A., Danalatos, G., Pasley, H., Pederson, C., Lamkey, K.R., 2020. Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt. Crop Sci. 60, 721–738. https://doi.org/10.1002/csc2.20039

Argentine Undersecretary of Agriculture, 2020. Estimaciones Agrícolas [WWW Document]. URL http://datosestimaciones.magyp.gob.ar/ (accessed 10.4.20).

Barbedo, J.G.A., 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Comput. Electron. Agric. 153, 46–53. https://doi.org/10.1016/j.compag.2018.08.013

Baum, M.E., Licht, M.A., Huber, I., Archontoulis, S. V., 2020. Impacts of climate change on the optimum planting date of different maize cultivars in the central US Corn Belt. Eur. J. Agron. 119, 126101. https://doi.org/10.1016/j.eja.2020.126101

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight Uncertainty in Neural Networks 37.

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. https://doi.org/10.1016/j.agrformet.2013.01.007

Buenos Aires Times | Corn to surpass soybean production this year, says Argentina [WWW Document], 2019. URL https://www.batimes.com.ar/news/economy/corn-to-surpass-soybean-production-this-year-says-argentina.phtml (accessed 12.5.22).

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Townsend, C., Solak, H., Semret, N., 2017. Crop yield predictions - high resolution statistical model for intra-season forecasts applied to corn in the US.

Campolo, J., Ortiz-Monasterio, I., Guerena, D., Lobell, D.B., 2022. Evaluating maize yield response to fertilizer and soil in Mexico using ground and satellite approaches. F. Crop. Res. 276, 108393. https://doi.org/10.1016/j.fcr.2021.108393

Cao, Z., Ma, L., Long, M., Wang, J., 2018. Partial adversarial domain adaptation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11212 LNCS, 139–155. https://doi.org/10.1007/978-3-030-01237-3_9

Chen, Junde, Chen, Jinxiu, Zhang, D., Sun, Y., Nanehkaran, Y.A., 2020. Using deep transfer learning for image-based plant disease identification. Comput. Electron. Agric. 173, 105393. https://doi.org/10.1016/j.compag.2020.105393

Chen, S., Liu, W., Feng, P., Ye, T., Ma, Y., Zhang, Z., 2022. Improving Spatial Disaggregation of Crop Yield by Incorporating Machine Learning with Multisource Data: A Case Study of Chinese Maize Yield. Remote Sens. 14, 2340. https://doi.org/10.3390/rs14102340

Chen, X., Feng, L., Yao, R., Wu, X., Sun, J., Gong, W., 2021. Prediction of maize yield at the city level in China using multi-source data. Remote Sens. 13, 1–17. https://doi.org/10.3390/rs13010146

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13. https://doi.org/10.1088/1748-9326/aae159

Cunha, R.L.D.F., Silva, B., 2020. Estimating Crop Yields with Remote Sensing and Deep Learning. 2020 IEEE Lat. Am. GRSS ISPRS Remote Sens. Conf. LAGIRS 2020 - Proc. 273–278. https://doi.org/10.1109/LAGIRS48042.2020.9165608

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. a J. R. Meteorol. Soc. 28, 2031–2064.

Deines, J.M., Patel, R., Liang, S.Z., Dado, W., Lobell, D.B., 2021. A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sens. Environ. 253, 112174. https://doi.org/10.1016/j.rse.2020.112174

Deodato, G., Ball, C., Zhang, X., 2019. Bayesian Neural Networks for Cellular Image Classification and Uncertainty Analysis. bioRxiv 824862. https://doi.org/10.1101/824862

EPA, U.S., 2001. United States Environmental Protection Agency. Qual. Assur. Guid. Doc. Qual. Assur. Proj. Plan PM Ambient Air 2.

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., 2020. Alfalfa Yield Prediction Using UAV-Based Hyperspectral Imagery and Ensemble Learning. Remote Sens. 12, 2028.

https://doi.org/10.3390/rs12122028

Feng, L., Zhang, Z., Ma, Y., Sun, Y., Du, Q., Williams, P., Drewry, J., Luck, B., 2021. Multitask Learning of Alfalfa Nutritive Value From 1–5.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci. data 2, 1–21.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian Active Learning with Image Data. https://doi.org/10.17863/CAM.11070

Ganin, Y., Lempitsky, V., 2014. Unsupervised domain adaptation by backpropagation. arXiv Prepr. arXiv1409.7495.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2017. Domain-adversarial training of neural networks. Adv. Comput. Vis. Pattern Recognit. 17, 189–209. https://doi.org/10.1007/978-3-319-58347-1_10

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 2030–2096.

Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266.

Gao, X., Huete, A.R., Ni, W., Miura, T., 2000. Optical-biophysical relationships of vegetation spectra without background contamination. Remote Sens. Environ. 74, 609–620. https://doi.org/10.1016/S0034-4257(00)00150-4

Gitelson, A.A., Viña, A., Ciganda, V., Rundquist, D.C., Arkebauer, T.J., 2005. Remote estimation of canopy chlorophyll content in crops. Geophys. Res. Lett. 32, 1–4.

https://doi.org/10.1029/2005GL022688

Global Yield Gap Atlas, 2021. Argentina - Global yield gap atlas [WWW Document]. URL https://www.yieldgap.org/argentina (accessed 11.15.21).

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Guan, K., Wu, J., Kimball, J.S., Anderson, M.C., Frolking, S., Li, B., Hain, C.R., Lobell, D.B., 2017. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. Remote Sens. Environ. 199, 333–349. https://doi.org/10.1016/j.rse.2017.06.043

H. Charles J. Godfray, John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, C.T., 2010. Food security: The challenge of feeding 9 billion people. Science (80-.). 327, 812–818. https://doi.org/10.1016/j.geoforum.2018.02.030

Härdle, W.K., Simar, L., 2019. Applied multivariate statistical analysis. Springer Nature.

Hodges, T., Botner, D., Sakamoto, C., Hays Haug, J., 1987. Using the CERES-Maize model to estimate production for the U.S. Cornbelt. Agric. For. Meteorol. 40, 293–303. https://doi.org/10.1016/0168-1923(87)90043-8

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int. Conf. Mach. Learn. ICML 2015 1, 448–456.

Jagtap, S.S., Jones, J.W., 2002. Adaptation and evaluation of the CROPGRO-soybean model to predict regional yield and production. Agric. Ecosyst. Environ. 93, 73–85.

https://doi.org/10.1016/S0167-8809(01)00358-9

Jensen, M.E., Burman, R.D., Allen, R.G., 1990. Evapotranspiration and irrigation water requirements. ASCE.

Jiang, H., Hu, H., Zhong, R., Xu, Jinfan, Xu, Jialu, Huang, J., Wang, S., Ying, Y., Lin, T., 2019. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. Glob. Chang. Biol. gcb.14885. https://doi.org/10.1111/gcb.14885

Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. Remote Sens. Environ. 112, 3833–3845. https://doi.org/10.1016/j.rse.2008.06.006

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. https://doi.org/10.1016/j.rse.2013.10.027

Kamir, E., Waldner, F., Hochman, Z., 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogramm. Remote Sens. 160, 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008

Kampe, T.U., 2010. NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. J. Appl. Remote Sens. 4, 043510. https://doi.org/10.1117/1.3361375

Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005

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, 064005. https://doi.org/10.1088/1748-9326/ab7df9

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? arXiv Prepr. arXiv1703.04977.

Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10, 1–10. https://doi.org/10.3389/fpls.2019.00621

Kouw, W.M., Loog, M., 2019. A review of domain adaptation without target labels. arXiv. https://doi.org/10.1109/tpami.2019.2945942

Kouw, W.M., Loog, M., 2018. An introduction to domain adaptation and transfer learning.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.

Li, Y., Guan, K., Schnitkey, G.D., DeLucia, E., Peng, B., 2019. Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. Glob. Chang. Biol. 25, 2325–2337. https://doi.org/10.1111/gcb.14628

Lin, C., Zhao, S., Meng, L., Chua, T.S., 2020. Multi-source domain adaptation for visual sentiment classification. AAAI 2020 - 34th AAAI Conf. Artif. Intell. 2661–2668. https://doi.org/10.1609/aaai.v34i03.5651

Liu, L., Xu, S., Tang, J., Guan, K., Griffis, T.J., Erickson, M.D., Frie, A.L., Jia, X., Kim, T., Miller, L.T., Peng, B., Wu, S., Yang, Y., Zhou, W., Kumar, V., Jin, Z., 2022. KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N2O emission using data from mesocosm experiments. Geosci. Model Dev. 15, 2839–2858. https://doi.org/10.5194/gmd-15-2839-2022

Liu, W., Ye, T., Jagermeyr, J., Müller, C., Chen, S., Liu, X., Shi, P., 2021. Future climate change significantly alters interannual wheat yield variability over half of harvested areas. Environ.

Res. Lett. 16. https://doi.org/10.1088/1748-9326/ac1fbb

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. https://doi.org/10.1146/annurev.environ.041008.093740

Lobell, D.B., Hammer, G.L., McLean, G., Messina, C., Roberts, M.J., Schlenker, W., 2013. The critical role of extreme heat for maize production in the United States. Nat. Clim. Chang. 3, 497–501. https://doi.org/10.1038/nclimate1832

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 U.S. Midwest. Science (80-.). 344, 516–519. https://doi.org/10.1126/science.1251423

Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks. 32nd Int. Conf. Mach. Learn. ICML 2015 1, 97–105.

Lu, Y., Nakicenovic, N., Visbeck, M., Stevance, A.-S., 2015. Policy: five priorities for the UN sustainable development goals. Nat. News 520, 432.

Luo, Z., Zou, Y., Hoffman, J., Li, F.F., 2017. Label efficient learning of transferable representations across domains and tasks. Adv. Neural Inf. Process. Syst. 2017-Decem, 165–177.

Ma, Y., Kang, Y., Ozdogan, M., Zhang, Z., 2019. County-level corn yield prediction using deep transfer learning, in: AGU Fall Meeting 2019. AGU.

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. https://doi.org/10.1109/TGRS.2023.3247343

Ma, Y., Zhang, Z., 2022. A Bayesian Domain Adversarial Neural Network for Corn Yield

Prediction. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Ma, Y., Zhang, Z., 2021. Multi-source Unsupervised Domain Adaptation on Corn Yield Prediction. AAAI Work. AI Agric. Food Syst.

Ma, Y., Zhang, Z., Kang, Y., Özdoğan, M., 2021a. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sens. Environ. 259, 112408.

Ma, Y., Zhang, Z., Lexie, H., Yang, Z., 2021b. An adaptive adversarial domain adaptation approach for corn yield prediction. Comput. Electron. Agric. 187, 106314. https://doi.org/10.1016/j.compag.2021.106314

Maaten, L. van der, Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237, 111599. https://doi.org/10.1016/j.rse.2019.111599

Malik, W., Boote, K.J., Hoogenboom, G., Cavero, J., Dechmi, F., 2018. Adapting the CROPGRO model to simulate alfalfa growth and yield. Agron. J. 110, 1777–1790. https://doi.org/10.2134/agronj2017.12.0680

Mehdipour Ghazi, M., Yanikoglu, B., Aptoula, E., 2017. Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing 235, 228–235. https://doi.org/10.1016/j.neucom.2017.01.018

Mishra, Prabhaker, Singh, U., Pandey, C.M., Mishra, Priyadarshni, Pandey, G., 2019. Application of student's t-test, analysis of variance, and covariance. Ann. Card. Anaesth. 22, 407.

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agric. For. Meteorol. 151, 385–393. https://doi.org/10.1016/j.agrformet.2010.11.012

Nasrabadi, N.M., 2007. Pattern recognition and machine learning. J. Electron. Imaging 16, 49901.

Omernik, J.M., 1987. Ecoregions of the conterminous United States. Ann. Assoc. Am. Geogr. 77, 118–125.

Omernik, J.M., Griffith, G.E., 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environ. Manage. 54, 1249–1266.

Park, S., Feddema, J.J., Egbert, S.L., 2005. MODIS land surface temperature composite data and their relationships with climatic water budget factors in the central Great Plains. Int. J. Remote Sens. 26, 1127–1144.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B., 2019. Moment matching for multi-source domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1406–1415.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G., 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. 7th Int. Conf. Learn. Represent. ICLR 2019 1–31.

Rodell, M., Houser, P.R., Jambor, U.E.A., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85, 381–394.

Rubel, F., Kottek, M., 2011. Comments on:" the thermal zones of the earth" by Wladimir Köppen (1884). Meteorol. Zeitschrift 20, 361.

Russell, S., Norvig, P., 2002. Artificial intelligence: a modern approach.

Russello, H., 2018. Convolutional Neural Networks for Crop Yield Prediction using Satellite Images. IBM Cent. Adv. Stud.

Saito, K., Watanabe, K., Ushiku, Y., Harada, T., 2018a. Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732.

Saito, K., Watanabe, K., Ushiku, Y., Harada, T., 2018b. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3723–3732. https://doi.org/10.1109/CVPR.2018.00392

Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2013. MODIS-based corn grain yield estimation model incorporating crop phenology information. Remote Sens. Environ. 131, 215–231. https://doi.org/10.1016/j.rse.2012.12.017

Schaaf, C., Wang, Z., 2015. MCD43A4 MODIS/Terra+ Aqua BRDF/Albedo Nadir BRDF Adjusted RefDaily L3 Global-500m V006. NASA EOSDIS L. Process. DAAC.

Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric. For. Meteorol. 284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Shen, J., Qu, Y., Zhang, W., Yu, Y., 2018. Wasserstein distance guided representation learning for domain adaptation. 32nd AAAI Conf. Artif. Intell. AAAI 2018 4058–4065.

Sheng, M., Liu, J., Zhu, A.X., Rossiter, D.G., Liu, H., Liu, Z., Zhu, L., 2019. Comparison of GLUE and DREAM for the estimation of cultivar parameters in the APSIM-maize model. Agric. For. Meteorol. 278, 107659. https://doi.org/10.1016/j.agrformet.2019.107659

Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G., Lobell, D.B., 2014. Testing remote sensing approaches for assessing yield variability among maize fields. Agron. J. 106, 24–32. https://doi.org/10.2134/agronj2013.0314

Soil Survey Staff, Natural Resources Conservation Service, USDA, 2020. Web Soil Survey [WWW Document]. URL https://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm (accessed 12.6.20).

Sun, C., Feng, L., Zhang, Z., Ma, Y., Crosby, T., Naber, M., Wang, Y., 2020. Prediction of End-Of-Season Tuber Yield and Tuber Set in Potatoes Using In-Season UAV-Based Hyperspectral Imagery and Machine Learning. Sensors 20, 5293.

Tasar, O., Tarabalka, Y., Giros, A., Alliez, P., Clerc, S., 2020. StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. 2020-June, 747–756. https://doi.org/10.1109/CVPRW50498.2020.00104

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, 2962–2971. https://doi.org/10.1109/CVPR.2017.316

USDA, 2020a. USDA National Agricultural Statistic Service (NASS) [WWW Document]. URL https://quickstats.nass.usda.gov/

USDA, 2020b. United States Department of Agriculture National Agricultural Statistics Service [WWW Document]. URL https://quickstats.nass.usda.gov/ (accessed 12.27.20).

Valentin Jospin, L., Buntine, W., Boussaid, F., Laga, H., Bennamoun, M., 2020. Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users. arXiv 1, 1–35.

Wagstaff, K.L., Lane, T., Roper, A., 2008. Multiple-instance regression with structured data.

Proc. - IEEE Int. Conf. Data Min. Work. ICDM Work. 2008 291–300. https://doi.org/10.1109/ICDMW.2008.31

Wang, A.X., Tran, C., Desai, N., Lobell, D., Ermon, S., 2018. Deep transfer learning for crop yield prediction with remote sensing data. Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Soc. COMPASS 2018. https://doi.org/10.1145/3209811.3212707

Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E.S., Li, H., 2018. Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2018-April, 4889–4893. https://doi.org/10.1109/ICASSP.2018.8461423

Wang, Y., Feng, L., Sun, W., Zhang, Z., Zhang, H., 2022. Exploring the potential of multi-source unsupervised domain adaptation in crop mapping using Sentinel-2 images. GIScience Remote Sens. 59, 2247–2265. https://doi.org/10.1080/15481603.2022.2156123

Wang, Y., Zhang, Z., Feng, L., Du, Q., Runge, T., 2020. Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. Remote Sens. 12, 1232.

Wang, Y., Zhang, Z., Feng, L., Ma, Y., Du, Q., 2021. A new attention-based CNN approach for crop mapping using time series Sentinel-2 images. Comput. Electron. Agric. 184, 106090. https://doi.org/10.1016/j.compag.2021.106090

Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L., 2018. Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3964–3973. https://doi.org/10.1109/CVPR.2018.00417

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 4, 3320–3328.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. 31st AAAI Conf. Artif. Intell. AAAI 2017 4559–4565.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716. https://doi.org/10.1016/j.rse.2020.111716

Yuan, W., Zheng, Y., Piao, S., Ciais, P., Lombardozzi, D., Wang, Y., Ryu, Y., Chen, G., Dong, W., Hu, Z., Jain, A.K., Jiang, C., Kato, E., Li, S., Lienert, S., Liu, S., Nabel, J.E.M.S., Qin, Z., Quine, T., Sitch, S., Smith, W.K., Wang, F., Wu, C., Xiao, Z., Yang, S., 2019. Increased atmospheric vapor pressure deficit reduces global vegetation growth. Sci. Adv. 5, 1–13. https://doi.org/10.1126/sciadv.aax1396

Zhang, Z., Jin, Y., Chen, B., Brown, P., 2019. California Almond Yield Prediction at the Orchard Level With a Machine Learning Approach. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.00809

Zhao, H., Zhang, S., Wu, G., Costeira, J.P., Moura, J.M.F., Gordon, G.J., 2018. Adversarial multiple source domain adaptation. Adv. Neural Inf. Process. Syst. 2018-Decem, 8559–8570.

Zhao, S., Li, B., Reed, C., Xu, P., Keutzer, K., 2020. Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv.

Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., Keutzer, K., 2019. Multi-source distilling domain adaptation. arXiv. https://doi.org/10.1609/aaai.v34i07.6997

Zhou, J., Wang, B., Fan, J., Ma, Y., Wang, Y., Zhang, Z., 2022. A Systematic Study of

Estimating Potato N Concentrations Using UAV-Based Hyper- and Multi-Spectral Imagery. Agronomy 12, 1–16. https://doi.org/10.3390/agronomy12102533

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2019. A Comprehensive Survey on Transfer Learning 1–27.

Zipper, S.C., Qiu, J., Kucharik, C.J., 2016. Drought effects on US maize and soybean production: Spatiotemporal patterns and historical changes. Environ. Res. Lett. 11. https://doi.org/10.1088/1748-9326/11/9/094021

Zuo, Y., Yao, H., Xu, C., 2021. Attention-based multi-source domain adaptation. IEEE Trans. Image Process. 30, 3793–3803. https://doi.org/10.1109/TIP.2021.3065254