2024

Developing Machine Learning Models for Selection of Management Zones

Sravanthi Bachina
South Dakota State University, Sravanthibachina60@gmail.com

DEVELOPING MACHINE LEARNING MODELS FOR SELECTION OF MANAGEMENT ZONES

BY

SRAVANTHI BACHINA

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Plant Science

South Dakota State University

2024

THESIS ACCEPTANCE PAGE
Sravanthi Bachina

This thesis is approved as a creditable and independent investigation by a candidate for the master's degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Kristopher Osterloh
Advisor Date

David Wright
Department Head Date

Nicole Lounsbery, PhD
Director, Graduate School Date

This dissertation is dedicated to the corner stones of my life: my Mom (Anu Radha Bachina) and Dad (Hanumantha Rao Bachina) whose unwavering support and guidance have shaped my journey; my husband Srinadh Kodali who had been my constant source of strength and encouragement; and to my loving son Sriansh Kodali who inspires me to do great things every day. Their love and sacrifices have been the wind beneath my wings, allowing me to soar to new academic heights.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to Dr. Kristopher Osterloh for his unwavering support, insightful guidance, and invaluable feedback throughout the journey of this thesis. His dedication and encouragement have been instrumental in shaping my research endeavors.

I am sincerely thankful to the committee members Dr. Jiyul Chang and Dr. Maitiniyazi Maimaitijiang for their expertise, constructive criticism, and commitment to excellence. Their collective wisdom has significantly enriched the quality of this work.

I am deeply grateful to NRCS (Natural Resources Conservation Service) and SDSU (South Dakota State University) for their generous funding and support, which made this research possible. Their financial assistance not only facilitated the execution of this study but also underscored their commitment to advancing scientific inquiry and promoting academic excellence.

I extend my heartfelt appreciation to both NRCS and SDSU for their investment in my academic pursuits, which have undoubtedly contributed to the development of knowledge in this field.

TABLE OF CONTENTS

ABBREVIATIONS

CEC cation exchange capacity

CM confusion matrix

CNN convolutional neural networks

DEM digital elevation model/s

EC electrical conductivity

KA kappa score

LiDAR light detection and ranging

MSE mean squared error

MZ management zone

NDVI normalized difference vegetative index

NIR near infrared

OA overall accuracy

PA producer accuracy

R^2 R-squared values

RF random forest

RMSE Root mean square error

SAR synthetic aperture radar

SH soil hydrology

SOM soil organic matter

SSURGO soil survey geographic database

SVM support vector machine

SWIR short wave infrared

TIFF Tag image file format

UA user accuracy

USA United States of America

VIP Variable importance projection

WHC water holding capacity.

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DEVELOPING MACHINE LEARNING MODELS FOR OPTIMAL SELECTION OF SOIL SAMPLING SITES

SRAVANTHI BACHINA

2024

Soil sampling and analyses play a crucial role in optimizing nutrient management and enhancing crop productivity. However, collecting representative samples across diverse landscapes is challenging due to knowledge gaps about spatial variability of soil properties, large fields, multiple samples, and analysis costs. Collecting soil samples based on the management zones can help farmers gather precise information about soil properties with fewer samples. Recent developments in precision agriculture and machine learning. This study aimed to develop machine learning models that can learn, analyze, and refine landscape and soil properties data for automated selection of soil sampling zones and generating prediction maps. Accordingly, random forest regression and classification models were built using data from four individual fields each with 12 features and five management zones for each field as target for model training and testing. Later a generalized model was developed by combining data from seven corn fields and seven soybean fields to improve predictive performance of the model which was evaluated on two new fields. The classification model for the four fields achieved overall accuracies of 0.71, 0.61, 0.75 and 0.69, kappa scores of 0.69, 0.58, 0.65 and 0.6, and F-scores of 0.7, 0.58, 0.75 and 0.59, respectively. Regression model yielded R^2 values of 0.71, 0.67, 0.83 and 0.76 and RMSE values of 6.7, 7.94, 5.45 and 2.4, respectively. The generalized model achieved overall accuracy of 0.8 and 0.75,

Kappa score of 0.71 and 0.59, F-1 score of 0.93 and 0.95 for soybean and corn field respectively. Despite achieving higher results generalized models failed to predict the management zones accurately. This could be due to limitation of model transferability and adaptability to various field conditions. This demonstrates the need to create and utilize high-resolution data with more spatial variability which will provide a comprehensive dataset for model training. Addition of other features including environmental variables, biomass indices that are more correlated to yield helps to improving the predictive performance of the models. Overall, this work establishes a foundational framework for novel applications of remote sensing data and machine learning techniques in addressing soil sampling challenges.

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Soil is an essential component of life on Earth, serving as a major interface between agriculture and the environment (Doran, 2002), sustaining human population and providing ecosystem services (Meyer & Turner, 2003). Understanding soil composition and its dynamic properties contributed to substantial developments in modern agriculture (Sengupta & Banerjee, 2012) by increasing crop productivity and farm profitability across the globe (Lal, 2008).

Soil sampling is the underlying tool aiding in a deeper understanding of soil properties (Carter & Gregorich, 2007) like nutrient availability, pH, sodium absorption ratio, cation exchange capacity (CEC), electrical conductivity (EC) (Corwin & Lesch, 2005). Although started as a scientific endeavor, by early 1900s both farmers and researchers began collecting soil samples and testing them for key plant nutrients (Peck, 1990). By analyzing the collected samples, researchers were able to gain a profound understanding of nutrient movement, uptake, and losses (Havlin, 2020; Sparrow et al., 1999; Wolf, 1999), while producers benefited from the resulting increase in crop productivity. Nationwide soil surveying in late 20th century further solidified the importance of soil sampling (Brevik et al., 2016) among masses.

Although sampling techniques have not changed significantly over the past decades (Knowles & Dawson, 2018), the number and quantity of soil samples required for various analyses have increased (D. W. Franzen & Peck, 1995). Modern agricultural machinery using precision agriculture practices like variable rate nutrient application require detailed information about different soil properties. As a result, the need for

collecting soil samples at regular intervals has intensified, with a shrink in the size of mapping units (Auernhammer, 2001; Kerry et al., 2010).

1.2 Soil testing

Within field variations of soil properties can be attributed to natural soil formation processes, environmental variation, and management practices. Factors like landscape position and soil parent material significantly impact soil texture, organic matter content, drainage, and other properties (Mallarino, 2023) directly influencing plant nutrient availability. Management practices affect nutrient levels by crop removal (Mallarino, 2023) can also be observed, thus requiring periodic soil testing. Modern agricultural practices like high density planting and nutrient demanding crops like wheat and corn pulls out more nutrients from soil necessitating nutrient addition for subsequent crops.

Soil fertility refers to the ability of soil to provide adequate amounts of essential nutrients for plant growth, thus directly influencing yields (Havlin, 2020). Fertilizers play a key role in maintaining soil fertility by replenishing nutrients that are depleted through crop uptake, erosion, leaching, anthropological and other natural processes. Under application of fertilizers reduces yields, thus directly impacting farm profitability (Penuelas et al., 2023). Over application on the other hand leads to runoff and leaching that contribute to eutrophication, algal blooms, and other environmental problems apart from the economic losses due to increased input costs (Savci, 2012). Therefore, testing soil samples for nutrients is essential for informed fertilizer applications, thus minimizing economic losses and mitigating environmental risks associated with over application.

Soil testing is crucial for assessing plant nutrient needs as it provides insights into the nutrient availability in soil. Apart from this, sampling plays a significant role in the process of soil testing and is often the primary source of error. Therefore, it is vital to collect soil samples that accurately reflect soil properties in the area of interest (Mallarino, 2005). To ensure an effective soil-testing program, several factors (like texture, organic carbon, water content, bulk density, EC etc.) with spatial and temporal variability must be considered as they influence nutrient concentrations in soils (Mallarino, 2005; Tomaz et al., 2022). Tomaz (2022) identified some of the important factors (EC, pH and nitrogen) influencing special and temporal variability of soil texture and chemical composition. Wang (2012) discussed the importance of spatial heterogeneity, spatial and temporal variability while designing sampling protocols. Therefore, it is crucial that sampling protocols address the wide range of variability in agricultural fields along with avoiding errors during sampling.

1.3 Representative samples

Accuracy and reliability of results from a sample depends on how precisely the sample represents an area of interest. Obtaining truly representative samples can be challenging (Tan, 2005; Wang et al., 2012). If the soil sample analyzed is not representative, the soil analysis results may not reflect the properties of the field in question (An et al., 2018; Cline, 1944). Instead, they can only describe the specific characteristics of the sample that was analyzed, leading to inaccurate or misleading information about the soil's nutrient levels, pH, and other important parameters (Tan, 2005). To ensure accurate soil analysis results, it is important to collect representative

samples from multiple optimal locations within the field or area being tested (Nawar & Mouazen, 2018; Wang et al., 2012).

1.4 Sampling techniques and challenges.

Soil sampling is a vital practice across various disciplines, including geology, environmental science, archeology, and agriculture (Brevik et al., 2015). However, sampling quantity, type and techniques employed across different disciplines are tailored to achieve specific objectives pertaining to those fields. For instance, in geology, soil samples are collected before exploring deeper layers, therefore, trenching, rock chip sampling, and coring techniques are preferred (Pennock et al., 2008). Grab sampling is popular in environmental sciences to assess contamination levels and soil quality from a relatively small sample (Mayer et al., 2014). In agriculture, grid sampling, zone sampling, random sampling, composite sampling, and depth specific sampling are employed to obtain soil samples based on the analysis that needs to be performed (Tan, 2005).

Random sampling is characterized by its simplicity and unbiased sample collection from random locations without following a predetermined pattern. In this method each soil core is selected randomly and independently from previously drawn units (Dinkins et al., 2008). Random sampling can be broadly categorized into a) simple random sampling, b) stratified random sampling (Cochran, 1977), c) systematic sampling (Madow & Madow, 1944). These sampling methodologies can be briefly described as random selection of sampling points throughout the fields, division of field into relatively homogenous areas before random sampling, and the use of grid or pattern for choosing sampling locations respectively. Since these methodologies doesn’t require precise selection of sampling locations drawing random samples is simple and straightforward to

implement (Dinkins et al., 2008). However, the heterogeneity of soil properties across the field is disregarded. While random sampling approach may be suitable for homogeneous fields, it fails to accurately represent the spatial variability across the field (Flint & Flint, 2002). Apart from this stratified random sampling needs additional data and expert knowledge to validate the results.

Composite sampling involves combining multiple soil samples taken from different locations within the field to create a representative composite sample, thus reducing both cost and time. Individual cores are collected in a diagonal or zig-zag pattern and combined to produce a composite sample (Lawrence et al., 2020). However, the number of samples required to get reliable results varies based on field conditions and the type of analysis performed (Lawrence et al., 2020). Past research provided recommendations for the number of soil cores necessary to create a representative composite sample, with the typical goal to estimate field-scale averages. Between 10 - 20 soil cores were frequently recommended for a field size of 10 - 20 acres (Hemingway, 1955; Tisdale et al., 1985), although this varied by nutrient, with nitrate requiring many more cores if higher precision was desired (45 cores for an 80% confidence interval within 10% of the mean on 75% of sampled areas) (Meisinger, 2015). Moreover zig-zag pattern of collecting soil cores is inadequate for log-normally distributed variables since there is no sufficient randomization (Lawrence et al., 2020) . Although useful to gain a general understanding of field conditions, this method of sampling fail to capture spatial and temporal variability of soil characteristics and requires more cores (Boswell & Patil, 1987) when using precision agricultural technologies.

Depth-specific sampling involves collecting soil samples at specific depths within the soil profile, typically at intervals such as 0-15 cm, 15-30 cm, and 30-60 cm. This method provides insights into nutrient distribution at multiple soil depths, crucial for understanding root development and nutrient uptake patterns. Additionally, depth-specific sampling helps identify nutrient stratification issues, where certain nutrients are accumulated or become depleted at specific depths (Reeves & Liebig, 2016). However, it requires additional effort and resources to collect and analyze samples from multiple depths and do not fully capture the variability present within each depth increment, particularly in soils with complex profiles. This method of sampling is less popular for sample collection among agricultural producers as it requires more resources and limited use cases.

Grid sampling involves dividing a field into a grid pattern and collecting samples within each grid cell or grid point (Clay et al., 2019; Knowles & Dawson, 2018). This method provides a systematic approach to assess soil variability across the field and obtain decent representative samples (Flint & Flint, 2002). Collecting samples at predetermined points allows for precise targeting of sampling locations and nutrient management interventions. Assigning the collection of soil samples to a third party is also simpler and easier with this method. Previous studies have used different grid sizes ranging from one acre grid to three to four-acre grids. The use of smaller grid sizes would provide more detailed soils maps that are essential for precision application of fertilizers. However, this increases the number of samples required in an acre and would thus increase overall cost of analysis. Although effective in fields with relatively uniform soil properties, where spatial variations in nutrient levels can be accurately captured, it can be

resource-intensive in large fields and may not capture variability present at scales smaller than the grid cell size. Therefore, challenges in determining the optimal grid size along with the need for a larger number of samples and the costs associated with them make its adoption challenging (Flowers et al., 2005).

Zone sampling entails dividing a field into homogenous sub regions within a field called management zones (Khosla et al., 2008; Milne et al.,2012.) based on factors such as soil type, topography, historical yield data, and remote sensing information (D. Franzen et al., 2000; Mallarino & Wittry, 2004). Delineating management zones can be challenging due to the complex relations and spatial variation of soil properties that affect crop yields (Park & Vlek, 2002). Although there are many methods to delineate management zones, natural breaks used by ArcGIS and ArcView software, diffuse conglomerate procedure used by FuzME (Minasny et al., 2007), and management zone analyst are the most common methods (Gili et al., 2017). The success or failure of these criteria depends on the objectives of management strategies. However, methods that incorporate spatial soil data in combination with k-means clustering procedures have been found to perform better in the studies conducted by Gili., (2017).

After delineating management zones, samples are collected randomly from each zone to estimate the characteristics of individual zones (Shaner et al., 2008). This method tailors sampling and nutrient management strategies to specific field conditions and management zones, allowing implementation of site-specific management practices. With potential benefits ranging from improving profitability to reducing environmental impacts, zones sampling is favorable over other sampling methods. Zone sampling is also preferable over other methods due to its resource use efficiency by focusing sampling

efforts where they are most needed. However, accurate delineation of management zones is crucial for the success of this approach and requires detailed soil surveys or advanced mapping techniques. Additionally, zone sampling may overlook variability present within individual zones, particularly if they are large or heterogeneous, necessitating careful consideration of zone boundaries and sampling density to ensure representative results.

Choosing the most suitable soil sampling technique can be challenging for producers, as each method comes with its own set of merits and drawbacks. Random sampling offers simplicity and ease of implementation, providing a broad overview of soil variability across the field. However, it is difficult to capture localized variations, leading to less representative results. On the other hand, grid sampling offers a systematic approach to capturing spatial variability, allowing for precise targeting of sampling locations. However, it is resource-intensive, especially in large fields, and may not capture variability at smaller scales. Zone sampling, meanwhile, tailors sampling efforts to specific field conditions and management zones, offering efficiency and site-specific management practices. However, appropriate delineation of management zones is crucial, and the method may overlook variability within individual zones if they are large or heterogeneous.

1.5. Soil heterogeneity

Beyond the choice of sampling technique, spatial distribution and variability of soil properties can pose additional challenges for producers (Asare & Segarra, 2018). Without proper scientific knowledge about the spatial distribution of soil properties, producers struggle to determine the optimal sampling strategy and analysis methods (Wang et al., 2012). This leads to suboptimal nutrient management decisions, potentially

resulting in reduced crop yields and increased input costs. Therefore, it is important to develop a protocol for optimal selection of sampling locations encompassing the developments in machine learning models.

1.6 Topography and its relationship with dynamic soil properties

Soil properties are inherently heterogeneous across relatively short distances (Webster, 2000). Therefore, comprehending spatial heterogeneity necessitates a substantial foundation of scientific knowledge and training, or an exhaustive compilation of historical soil datasets pertinent to the area under investigation (Wang et al., 2012). Distribution of physical and chemical soil properties is of significant interest due to their direct and indirect impact on productivity, particularly for site-specific fertility management.

Topography is one of the five fundamental elements of the soil forming factors (Jenny, 1941) that can significantly influence a wide array of soil physical and chemical properties (Miller & Schaetzl, 2015). Topography is often characterized by terrain attributes such as slope, aspect, and curvature, maintains intricate connections with a wide range of soil properties (Ceddia et al., 2009; Kumhálová et al., 2011). Ruhe and Walker's 1968 model introduced five major hillslope profile positions, widely applicable across different climates, landscape ages, and parent materials (Walker, 1968). This framework has been substantiated by subsequent research, becoming a standard reference for investigating soil variability across hillslopes and for broader landscape characterizations. However, these relationships extend beyond soil, as hillslope position serves as a valuable framework in ecological studies, linking soil and vegetation dynamics (Miller & Schaetzl, 2015).

Soil properties (physical, chemical, and biological) change significantly with topography (Ceddia et al., 2009) affecting soil water content and crop responses (Ayele et al., 2015). Understanding intricate relationships between topography and various soil properties such as texture, soil organic matter (SOM), cation exchange capacity (CEC), soil hydrology (SH), water holding capacity (WHC) and aggregate stability aid in making better management decisions (Kumhálová et al., 2011; Seibert et al., 2007). Thus, separating soil samples by landscape position can provide detailed information for precision agricultural practices by informing spatial distributions of soils data (Reza et al., 2015).

Soil texture is one the important soil physical properties influenced by topography. In areas with steep slopes, soil texture tends to be coarser due to higher erosion rates that remove finer particles leaving behind sandier soils. Conversely in low lying areas or depressions where water and finer sediments accumulate, soils are typically fine textured, with higher proportions of silt and clay (N. C. Brady & Weil, 2016). This variation in soil texture influenced by topography affects soil water retention, nutrient availability, and organic matter content, thereby influencing plant growth and yields (N. Brady & Weil, 2004).

The relationship between topography and aggregate stability in soils is well documented with multiple studies showing stability varying across the landscape. Pierson & Mulla, (1990) found that aggregate stability was higher in depressions and decreased towards summit positions, a trend driven by depletion of organic matter in higher elevations due to erosion. Jakšík, (2015) observed a similar pattern while studying the impact of terrain attributes (slope, curvature, and aspect) on aggregate stability. Overall,

variability in aggregate stability across a field is closely tied to material redistribution and organic matter content, influenced by terrain attributes.

Soil water holding capacity is another soil physical property influenced by topography. Previous studies have underscored the importance of topography in influencing WHC (H. Yang et al., 2021) and therefore being used to accurately predicting it (Fathololoumi et al., 2021; Obi et al., 2014). Shape and slope of the land affect runoff, infiltration, and depth of soil layers which in turn influence the amount of water retained in soil (H. Yang et al., 2021). Steeper slopes make soil more prone to runoff, reducing infiltration and consequently the WHC of soils. Conversely flat areas with a gentler slope promote water retention as they allow more water to infiltrate through, soil increasing WHC. Furthermore, topographical features lead to the formation of microclimates and variation in soil composition across different landscapes, affecting the distribution of moisture and WHC of soil (H. Yang et al., 2021).

Spatial heterogeneity of SOM can be better understood by considering topographic features like elevation and aspect (Zhu et al., 2019). Aspect influences spatial patterns of SOM through altering solar radiation on hill slopes (Lybrand & Rasmussen, 2015). However, this trend can only be valid with greater changes in elevation across the landscape (Chen et al., 2016). In places with gentle landscapes across a wide area, elevation indirectly influences SOM by altering the physical soil properties such as soil texture, aggregate stability and WHC as discussed earlier.

The complexity involved in collecting a representative sample following scientific procedures can be challenging for producers (Oliver et al., 2010). Adding to this, the quantity of samples to be collected and the cost of analysis are pushing producers to cut

back on the quantity of soil samples collected and therefore quality of soils data (Flint & Flint, 2002). Results from such sampling cannot be reliably used with precision agriculture practices thus making them difficult for adoption. However, dividing soil sampling area by management zones and utilizing recent developments in precision agriculture and machine learning methods can provide a better alternative to the problem of selecting optimal soil sampling locations.

1.7 Management zones

Management zones (MZ) are specific areas within a field that are managed individually based on their distinctive characteristics like topography, yield data, remote sensing data, and other agronomic factors (Chang et al., 2003). Delineating management zones based on such factors has been a common practice in multiple studies involving site specific precision agricultural practices (Chang et al., 2003; Flowers et al., 2005; Nawar et al., 2017). However, delineation of management zones based on yield data offers advantages over other factors (Flowers et al., 2005). Yield monitoring systems have been in use with most of the producers over the past decades. Yield data from these systems can provide reliable information on the performance of various management factors influencing yield. Early research found that delineation based on multiyear yield data could be related to soils data (Lark & Stafford, 1997; Miao et al., 2018; Nawar et al., 2017). Subsequent research also suggested that using multiyear average yield maps to delineate management zones for soil sampling is promising (Blackmore, 2000; Diker et al., 2004; Lark & Stafford, 1998). Therefore, using existing yield data/maps to delineate management zones can be an affordable alternative to developing maps based on complex interacting factors, but existing yield data doesn’t transfer well between crops.

1.8 Machine learning in soil science

Machine learning is a branch of artificial intelligence and computer science involving computational algorithms that are designed to emulate human intelligence (El Naqa & Murphy, 2015). With potential to analyze large and complex data sets, use of machine learning techniques has been on a raise across diverse fields in the past decade (Padarian et al., 2020). Pattern recognition, computer vision, engineering, finance, biological and biomedical sciences are a few areas where machine learning techniques have been successfully applied (El Naqa & Murphy, 2015).

In soil science machine learning applications have been used to develop models that analyze and estimate various soil parameters like moisture content (Ahmad et al., 2009), bulk density (Bondi et al., 2018), soil organic carbon, parent material and others (Padarian et al., 2020). Among other applications in soil science, prediction of soil types and properties using digital soil mapping and pedotransfer functions are prominent (A. McBratney et al., 2019; A. B. McBratney et al., 2003). Increasing availability of soil data from multiple remote sources alongside the developments in geographical information systems and availability of free open-source algorithms have led to increased adoption of machine learning techniques in soil science (Padarian et al., 2020).

Incorporating machine learning models can substantially simplify and enhance the process of delineating management zones with the potential to analyze complex datasets. Furthermore, multiyear yield data, soil characteristics, and remote sensing information can be analyzed to identify patterns and relationships that may not be immediately apparent (Pham et al., 2024). Machine learning algorithms also have the potential to predict effective management zones, tailor recommendations for each specific area, and

refine these suggestions based on new data. This automation reduces the need for intensive manual analysis and allows for dynamic adjustment of management zones with changing conditions.

1.9 Review of recent models in soil science

Adoption of machine learning techniques in soil science have led to the development of multiple models with tailor made algorithms (Padarian et al., 2020) like multi objective optimization (Kazemi & Samavati, 2023), maxvol, (Petrovskaia et al., 2021) convolutional neural networks (CNN) (Pham et al., 2024) etc. These models have significantly enhanced our ability to predict soil properties, understand soil composition, and optimize agricultural practices. Although dealing with a wide variety of problems, the overall approach on developing and deploying these models have been consistent across multiple studies (Hengl et al., 2003; Kazemi & Samavati, 2023; Petrovskaia et al., 2021). However, some models prioritize specific soil parameters (Ahmad et al., 2009; Bondi et al., 2018; Padarian et al., 2020) while others emphasize on scalability for large geographical areas (Kazemi & Samavati, 2023). Additionally, differences in incorporating remote sensing data, landscape data, and approaches in delineating management zones highlight the diverse research being done in this area.

Using multi objective optimization algorithm, Kazemi & Samavati (2023) were able to determine optimal sites for soil sampling. This model has successfully integrated remote sensing data (NDVI -normalized difference vegetative index) and digital elevation model (DEM) to delineate management zones. Maxvol's algorithm utilizes the concept of D-optimal design, seeking to select sample locations with the most significant dissimilarities in topographical features (slope, aspect, topographic wetness index, closed

depressions). This aims to capture the variability of soil cover with a considerably smaller dispersion in prediction error compared to existing approaches for spatial soil sampling. On the other hand, CNN models utilize a deep learning model with encoder-decoder architecture, self-attention mechanism and atrous convolutional networks to process input data like slope, aspect, flow accumulation, NDVI and yield data in predicting optimal soil sampling sites (Pham et al., 2024).

Most published models rely heavily on topographical characteristics as predictors, potentially overlooking other properties that could provide valuable information for selecting sampling sites, such as soil properties and satellite imagery with different spectral, temporal, radiometric, and spatial resolutions. Moreover, these studies also fail to capture remote sensing parameters like short wave infrared band (SWIR), synthetic aperture radar bands (Domenech et al., 2020), near infra-red bands (NIR) (Ahmadi et al., 2021) that are corelated to soil properties (Domenech et al., 2020). Remote sensing techniques also have potential for identification and quantification of different soil properties that exhibit unique spectral signatures (Abdulraheem et al., 2023). Reflectance patterns observed in different wavelength ranges can provide information about soil properties. NIR is sensitive to soil moisture content and clay minerology, whereas SWIR can be used to estimate soil organic carbon content, (Balaram and Sawant, 2022). Additionally, radar sensors can provide information about soil moisture content, surface roughness and texture (Petropoulos et al., 2015). Therefore, integrating remote sensing parameters alongside topographical characteristics offers a more comprehensive approach for selecting sampling sites.

Random forest (RF) is often considered more useful compared to multi objective optimization, Maxvol, and CNN models due to its robustness, versatility, and performance in various applications. According to Motia & Reddy, (2021) RF is one of the most commonly used machine learning techniques in soil science and agriculture domain. The RF model is useful to perform both classification and regression tasks. RF model is a versatile ensemble machine learning algorithm comprises multiple decision trees, each built on a random subset of the training data (bootstrapping) and a random subset of features (feature randomization) (Breiman, 2001). This ensemble approach mitigates overfitting and increases predictive accuracy. During training, each tree independently makes predictions, and final prediction is obtained through majority voting (for classification) or averaging (for regression). RF model also handles categorical data, like hillslope position or geology type, which most other models cannot. RF excels in reducing overfitting, delivering high predictive accuracy, and providing feature importance insights. For this study random forest was used to perform classification and regression tasks.

When developing optimal soil sampling locations using machine learning approaches, literature (Kazemi & Samavati, 2023; Petrovskaia et al., 2021; Pham et al., 2024) has revealed a variety of techniques that have been tailored to address specific aspects of soil science. Prevailing models, such as multi objective optimization, Maxvol, and CNN, have each contributed to the prediction of optimal soil sampling locations. Despite their innovative approaches, these models have often emphasized specific topographical features, occasionally at the expense of integrating a more holistic view that encompasses a variety of remote sensing data, landscape characteristics, and intrinsic

soil properties. This oversight presents a gap that our study aims to address by leveraging the robustness and versatility of random forest model.

In this study we aim to develop a comprehensive model that integrates a wide spectrum of data, including high-resolution remote sensing imagery, comprehensive landscape data, and detailed soil properties. Incorporation of diverse datasets promises to enhance the model's sensitivity to the multifaceted nature of data thus providing a more nuanced understanding of spatial variation of soil properties across varying landscapes.

Building upon previous literature highlighting the limitations of solely relying on topographical characteristics, our research seeks to achieve several objectives. The primary objective of this study is to collect readily available data encompassing soil properties, landscape features, and remote sensing imagery for the development of random forest models. By harnessing the strengths of Random Forest algorithms, we intend to construct a more integrative model that can learn, analyze, and refine the data for automated and reliable selection of soil sampling zones.

The overarching goal of this project is to develop a sophisticated tool capable of capturing spatial variations in soil properties across diverse landscapes, thereby facilitating more informed decision-making in soil sampling efforts. Through integrating multidimensional datasets and utilizing machine learning techniques, we hypothesize that our model would aid in optimal selection of soil sampling locations in reliable and cost-effective manner. Ultimately, our study aims to provide a practical solution to the challenge of selecting optimal soil sampling sites by leveraging the full spectrum of available data and analytical tools.

CHAPTER 2: MATERIALS AND METHODS

2.1 Data Collection and Preprocessing

2.1.1 Study Sites

The data used in this study were collected from two counties in South Dakota, namely Aurora and Davison County (Figure 1), and encompassed a total of 18 field sites, each field covers approximately an average of 220 acres area that are mostly used for cultivation of corn and soybean. Specific site locations were redacted to protect producer privacy. These field sites are managed by local producers who provide access to field boundaries, management, and yield data. All the fields are relatively flat (1-3% slope), glaciated till plain terrain.

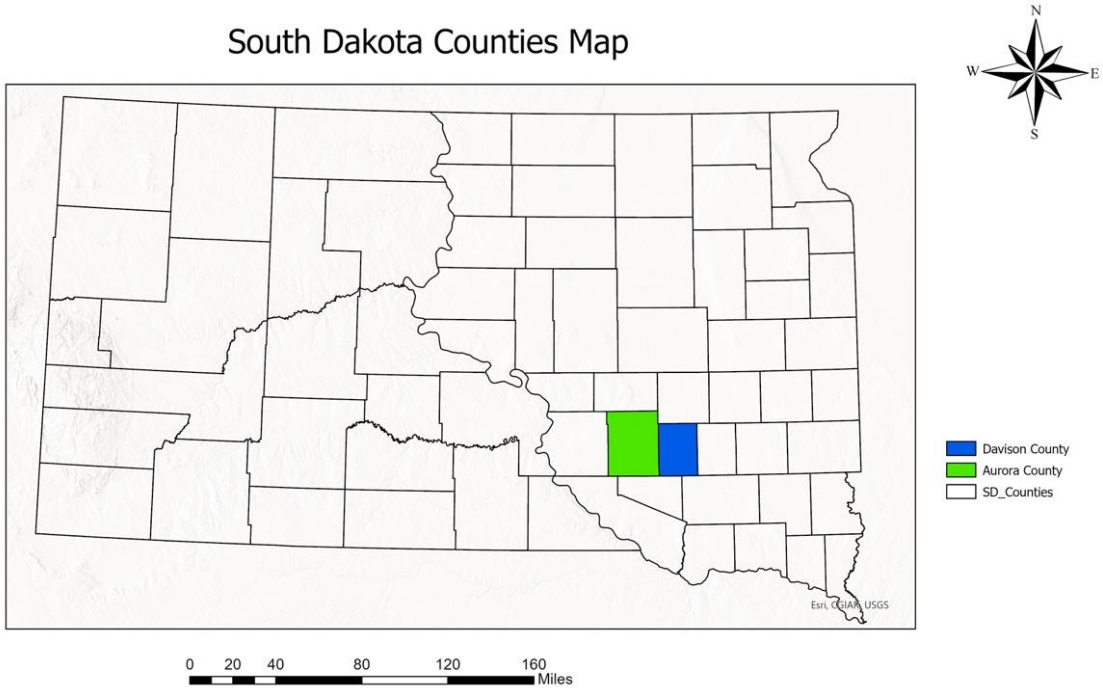

Figure 1. Map of South Dakota highlighting Aurora and Davison counties.

2.1.2 Data Collection

The objective of this study is to develop a machine learning model that can learn, analyze, and refine landscape data, soil properties data and remote sensing data from readily available sources for automated generation of management zones. To derive landscape attributes (slope, aspect, and flow accumulation), digital elevation models (DEMs) for each field were acquired from LiDAR (Light Detection and Ranging) dataset in South Dakota Geological Survey (https://www.sdgs.usd.edu/) with a spatial resolution of one meter and were processed in ArcMap 10.8 version (Esri®, ArcGIS, ArcMap 10.8) using surface raster function. Soil properties data including water holding capacity (WHC), soil texture, cation exchange capacity (CEC), soil hydrology (SH), and soil organic carbon (SOC) were obtained from the Soil Survey Geographic Database (SSURGO) dataset (*Soil Survey Geographic Database (SSURGO) | Natural Resources Conservation Service*, 2022) .

Landscape attributes and soil properties data thus obtained were then selected as features for training and testing the RF models. Results obtained from using landscape and soil properties had very low accuracy levels and the models had difficulty in predicting management zones. Variable importance projection plots developed during training and testing of RF models showed that soil properties data minimally contributed to the predictive performance of models. Soil properties data obtained from SSURGO data were of low resolution contributing from low to no variability across the fields, so no utility in using it to subdivide the fields. Therefore, given the lower resolution of the initial soil properties data, augmentation was performed by using remote sensing data sourced from Sentinel-2A satellite imagery from Copernicus Open Hub website with a

spatial resolution of 20 meters. Raw spectral bands such as NIR, and SWIR are obtained on June 14 in the year 2022. SAR-C bands were obtained from sentinel 1A on June 14, 2022, with a spatial resolution of 20 meters. These remote sensing data sources were utilized to enhance the precision and granularity of soil information due to their inherent correlation with soil properties (Balaram & Sawant, 2022; Petropoulos et al., 2015).

2.1.3 Creation of management zones

Management zones delineated from multiyear average yield data were utilized as target or dependent variables. Three year's crop harvests yield data was obtained from producers, this yield data was collected based on the crop type. A multiyear average yield map was then generated by using multiyear average yield analysis wizard in SMS Ag software (Ag Leader®, SMS Advanced). Later this multiyear average yield map was divided into 5 management zones by using natural breaks in SMS Ag software. These zones were labeled as 1, 2, 3, 4, and 5 ranging from low yielding zones to high yielding zones respectively.

2.1.4 Data Preprocessing

All the features and target data layers were saved as raster files in TIFF format. Subsequently, these raster files were then combined into a single raster stack. To facilitate model training and testing, a point shapefile was generated by using raster to point conversion tool in Arc Map software for each field. The point shape files were then used to extract data from the stacked raster, creating data frames which had all the features (8 features-before using remote sensing data and 13 features-after using remote sensing data) and target data needed to be used for subsequent model training and testing. The entire process of data extraction through point shape file was done by using the

"*Extract multi values to point*" tool in ArcMap 10.8 version (Esri®, ArcGIS, ArcMap 10.8).

2.2 Model development

2.2.1 Classification of management zones

Classification of a field into different management zones depends on a plethora of factors ranging from the distribution of soil properties to target management practices. Therefore, the number of management zones are selected based on the practicality of the target management operation. To offer flexibility in adoption we generated five management zones for each field.

Two scenarios were used for classification of management zones. One is developing a classifier model for categorical prediction of 5 management zones and in scenario two a regressor model is used to predict the multiyear average yield values (continuous variables) (Table 1) which are classified later by using "*natural breaks*" of classification function in ArcMap 10.8 version (Esri®, ArcGIS, ArcMap 10.8).

Table 1. Multiyear average yield statistics for all fields in the study.

Fields*	Min yield (bu/ac)	Max yield (bu/ac)	Mean yield (bu/ac)	sd**	Crop type
PO	34.24	69.1	50.86	5.77	Soybean
MU	29.38	53.65	43.58	3.5	Soybean
LW	26.87	57.75	48.83	3.84	Soybean
LN	15.55	58.41	48.13	4.88	Soybean
LH	22.34	61.39	48.97	4.42	Soybean
KH	15.00	50.75	39.2	4.45	Soybean
HI	25.61	46.45	37.89	4.4	Soybean
CW	17.78	59.42	43.1	4.94	Soybean
AD	23.45	73.18	60.74	6.72	Soybean
MU	110.27	188.8	156.23	11.41	Corn
LS	41.88	197.67	147.66	25.01	Corn
KS	83.19	193.3	151.81	13.63	Corn
KM	47.15	175.96	140.88	12.96	Corn
KE	109.45	172.87	149.25	11.35	Corn
HT	54.98	144.44	107.23	14.17	Corn
FN	68.27	195.97	134.71	22.21	Corn
BS	101.87	169.02	146.03	13.25	Corn
AT	121.5	171.82	148.27	10.33	Corn

* Field names were assigned based on a two-letter format used by the producers.
** Standard deviation

2.2.2 Data Splitting

To facilitate model training and testing, the dataset was divided into two subsets: a training set and a testing set. This partitioning was carried out to ensure the model's ability to generalize its predictions beyond the training data. The data was split into training and testing data sets in 70: 30 ratios randomly for each field. The training set, comprising approximately 70% of the data, was designated for model training. The model learns from this portion of the data to make predictions. The remaining data points, approximately 30% of the dataset, constituted the testing set. This set remained

untouched during the model training phase and was reserved solely for evaluating the model's performance. Testing set serves as an independent validation dataset to assess the model's ability to generalize unseen data.

Table 2 shows the total number of data points used for training and testing the models for each field.

Table 2. Total number of data points used for training and testing the models for selected fields.

Fields	Total data points - Training	Total data points - Testing
KM	9289	3981
KH	6768	2901
BS	16339	7003
HT	8161	3497
Combined Soybean	27243	11675
Combined Corn	33646	14301

2.2.4 Data Normalization and Hyperparameter tuning.

Most machine learning algorithms, including RF, SVM (Support Vector Machine), and others, are sensitive to the scale of input features. Features with larger scales can dominate those with smaller scales in the model learning process. Data normalization is a preprocessing step that scales or transforms features in the dataset into a standard range or distribution. The primary goal of data normalization is to ensure that different features have similar scales (García et al., 2015). Because of this importance we normalized the data used in this study using Min-max scaling process to make sure all input variables had the same scale in variability. This scaling process transforms the data so that it falls within a specific interval, typically [0,1].

Hyperparameters are settings or configurations that are not learned from the data but are set prior to training. Hyperparameter tuning is the process of finding the optimal hyperparameters for a machine learning model (Weerts et al., 2020). This tuning process serves several essential purposes including maximizing model performance (Diaz et al., 2017), preventing overfitting and enhancing the overall efficiency of the model (L. Yang & Shami, 2020). According to Padarian, (2020) many studies in soil science that have employed machine learning techniques have observed a significant improvement in their results when hyperparameter tuning was employed.

In this study hyperparameter tuning was carried out using R programming libraries, specifically "randomForest" for Random Forest models within the R Studio environment. For random forest regression model basic hyperparameters include feature sampling, number of trees and feature subset size. In feature sampling the number of features randomly sampled at each split when building decision trees are coded as “mtry” in R. The number of decision trees that will be built and aggregated were coded as “ntree”, and feature subset size coded as “max_features” in R. In RF classification model, hyperparameters were split quality (coded as “criterion”) and tree depth limit (coded as “max_depth”). Split quality includes options for measuring impurities (coded as "gini” and "entropy" in R). Upon tuning, best parameters (Table 3)were used to fit the model for training and testing (Posit team, 2023).

Table 3. Best parameters obtained after hyperparameter tuning.

Hyperparameters used	Best parameters
mtry	8
ntree	200
max_features	log2
max_depth	10
criterion	gini

2.3 Evaluation metrics

In the evaluation of classification models, a comprehensive set of scientific metrics was employed, including overall accuracy (OA), kappa scores (KA), and F-1 scores, alongside the use of confusion matrices (CM) to gain insights into model performance in predicting 5 management zones. For regression models, the assessment centered on R^2 values and Root Mean Square Error (RMSE), which measured model fit and predictive accuracy. Additionally, scatter plots were utilized to visually inspect the correspondence between predicted and actual multiyear average yield values. To further enhance the analytical depth, variable importance projection (VIP) plots were generated, these plots are visual representations that reveal the significance of different features in a machine learning model's prediction, playing a crucial role in both regression and classification models (Greenwell et al., 2020). In regression models, metrics like "percentage increase in mean squared error" (%IncMSE) and "increase in node purity" (IncNodePurity) are employed to gauge the importance of individual features. "%IncMSE" quantifies how much each feature influences the model's accuracy in predicting numerical values, while "IncNodePurity" assesses their role in improving node purity within decision trees, leading to more precise predictions. Higher values for these metrics signify greater feature importance, aiding in feature selection and model interpretation (González et al., 2015).

In classification models, "mean decrease Gini" is a vital metric within VIP plots, assessing each feature's contribution to reducing impurity and enhancing class separation. Features with higher "mean decrease Gini" scores are pivotal in effectively distinguishing between different classes, while lower scores indicate less influence (Han et al., 2016).

This metric assists in identifying key variables for accurate classification and informs model interpretation and feature selection. By ranking features based on their impact, these plots aid in model interpretation, feature selection, and identifying influential factors (Greenwell et al., 2020). They provide a clear and intuitive way to understand which variables contribute most to the model's performance allowing for the exploration of key features that significantly influenced model predictions. These rigorous evaluation methods, coupled with variable importance analysis, provided a robust and scientifically grounded basis for comparing and drawing conclusions about the efficacy and accuracy of the models.

Following a comprehensive evaluation, prediction maps with five management zones were generated for each field. Predicted maps were then visually compared to the actual maps to identify for any misclassifications. Specific sections of the fields that had some degree of misclassification were identified to gain insights into the extent of misclassification. This could potentially aid in identifying the zones in which the model struggled to make accurate predictions.

2.4 Workflow

In initial tests, to achieve the best accuracy levels different covariates were assessed. Initially, we selected four random fields (BS, KM, KH, and HT) and conducted model training and testing individually using only landscape and soil properties data. However, these efforts yielded suboptimal accuracy levels. To address this issue, we employed Variable Importance projection plots (VIP), which revealed that soil properties data had the lowest contribution to model performance. To enhance accuracy, we incorporated remote sensing data, including features like NDVI (Normalized Difference

Vegetation Index), NIR bands, SWIR bands, and SAR bands. These remote sensing data sources were chosen because of their known relationships with soil properties, backed by prior research (Abdulraheem et al., 2023; Balaram & Sawant, 2022; Domenech et al., 2020; Nawar & Mouazen, 2018; Petropoulos et al., 2015) which improved model performance.

To enhance the models' generalizability, we undertook further steps by incorporating a broader dataset. We combined data from seven corn fields (BS, HT, KE, KM, KS, LS, and MU) and seven soybean fields (AD, HI, LH, LN, LW, KH, and MU). Subsequently, separate models were trained and tested for each crop type. These models were then evaluated on two new fields for corn (FN and AT) and soybean (PO and CW) to assess model performance. This approach allowed us to ensure that the models were capable of generalizing their predictions to different field types, both for corn and soybean cultivation.

Throughout this iterative process, evaluation metrics, VIP Plots, and prediction maps at each stage were generated and are further elaborated in results and discussion sections. The overall workflow of the project was further illustrated in Figure 2

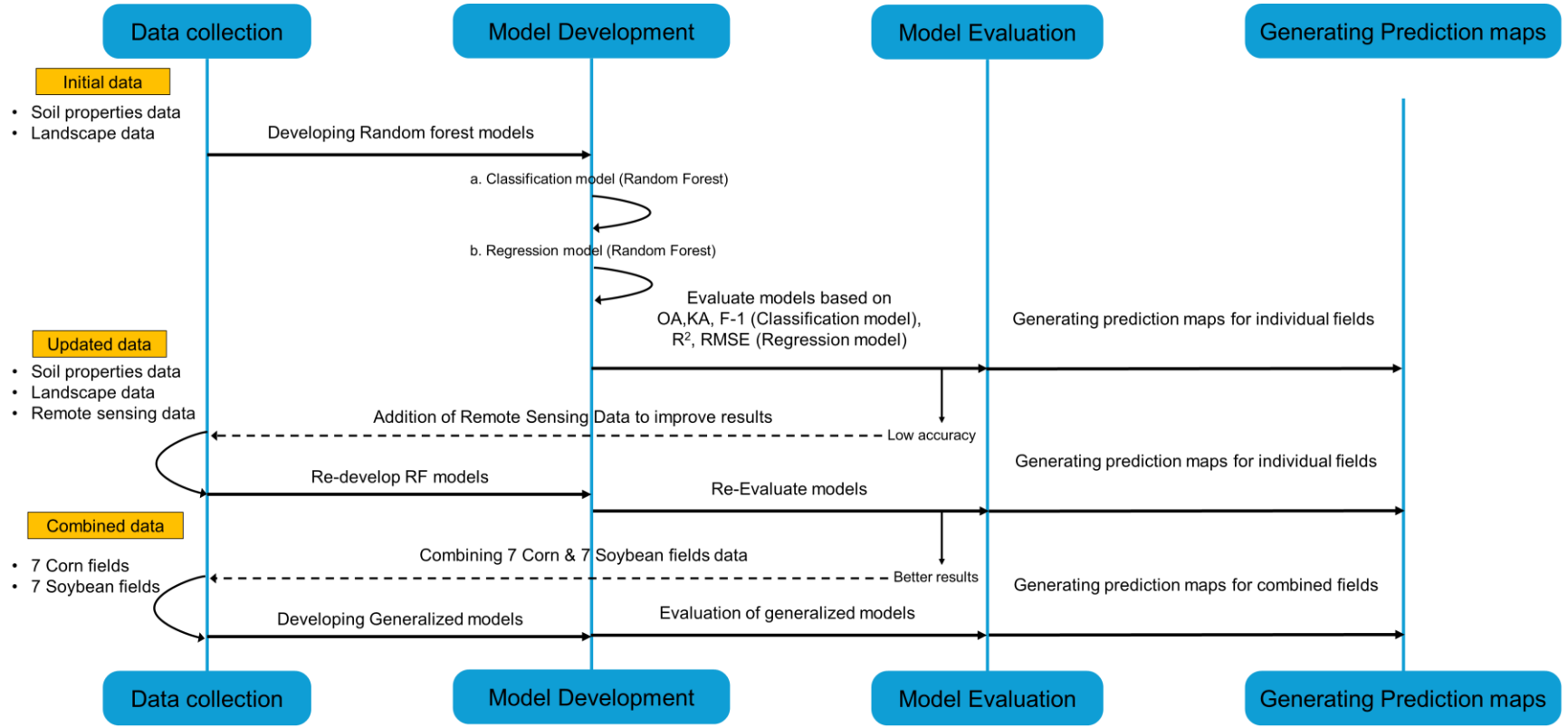

Figure 2. Project workflow diagram

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Classification Model.

Using the RF classification model, overall accuracy (OA) levels of four fields (HT, KM, KH, BS) before incorporating remote sensing data ranged from 0.38 to 0.49. Incorporating remote sensing data into the model improved the OA values by an average of 24% across the four fields. Kappa scores (KA) for the same fields ranged from 0.18 - 0.33 and increased on an average of 33% after including remote sensing data. Whereas F-1 scores showed a dissimilar pattern with scores ranging from 0.76 – 0.9, decreasing between 1-4 percent across all fields with an exception in BS field, while still indicating better precision and recall after incorporating remote sensing data (Figure 3).

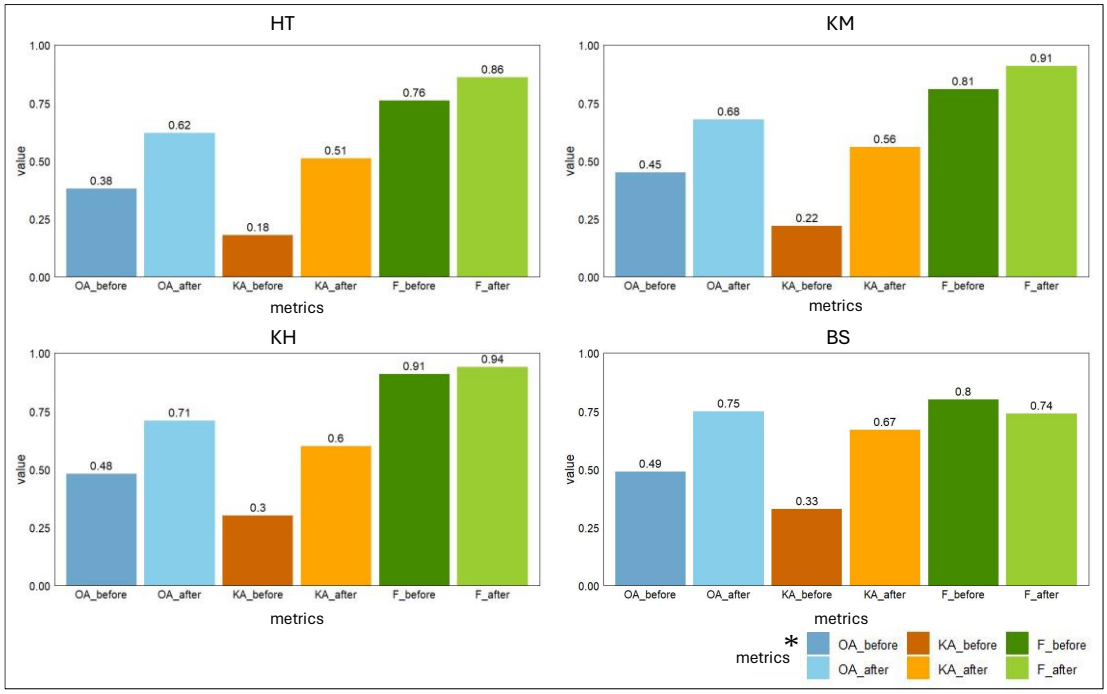

Figure 3. Comparison of accuracy levels of RF model before and after adding remote sensing data for HT, BS, KM and KH fields.

*(OA – Overall Accuracy, KA – Kappa Score, F – F-1 Score).

In addition to OA, KA and F-1 scores, CM's (Table 4) provide crucial insights into the performance of classification models using metrics like Producer Accuracies (PA) and User Accuracies (UA) for a specific zone within each field. PA and UA values aid in uncovering the model's ability to properly classify management zones while highlighting false positives and false negatives. Before the incorporation of remote sensing data, PA values ranged from 0.22 to 0.62 while UA values ranged from 0.1 to 0.62 across different zones in the four fields under investigation. A noteworthy transformation occurred in the RF model after the inclusion of remote sensing data where the PA and UA values increased by an average of 27% and 25% respectively. The increase in the accuracies can be attributed to using NIR, SWIR and SAR bands as features obtained from remote sensing data. This underscores the value of rich, multidimensional feature sets in enhancing model performance.

On the other hand, data obtained from SSURGO data sets (available for entire USA), exhibit very limited variability among soil properties when clipped to the individual field levels due to the variation in the scale. The lack of variation across the fields in initial data greatly affects the model performance as there are fewer variations across features to learn from. A more robust data set with spatially variable features exposes the model to a more comprehensive set of scenarios, enabling it to learn from more general patterns across the field. This broad exposure helps in fine tuning model's decision boundaries to handle complex, non-linear relationships more effectively, leading to more accurate predictions.

Table 4. Confusion matrices for RF classification model.

Before adding remote sensing data								After adding remote sensing data							
BS								**BS**							
Zones	**1**	**2**	**3**	**4**	**5**	**Total**	**UA****	**Zones**	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**
1	114	145	66	58	41	424	0.26	**1**	273	114	16	7	0	410	0.66
2	103	516	170	173	116	1078	0.47	**2**	54	819	136	43	14	1066	0.76
3	45	195	429	401	247	1317	0.32	**3**	15	138	822	255	96	1326	0.62
4	34	92	256	1144	595	2121	0.53	**4**	1	22	147	1599	312	2081	0.76
5	28	78	145	552	1315	2118	0.62	**5**	0	7	31	324	1758	2120	0.8
Total	324	1026	1066	2328	2314	7058		**Total**	343	1100	1152	2228	2180	7003	
PA**	0.35	0.5	0.4	0.49	0.56			**PA**	0.79	0.74	0.71	0.71	0.8		
KM								**KM**							
Zones	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**	**Zones**	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**
1	11	19	27	43	6	106	0.1	**1**	49	31	12	2	1	95	0.51
2	19	81	130	166	36	432	0.18	**2**	8	224	115	22	8	377	0.59
3	12	101	452	437	138	1140	0.39	**3**	7	72	707	284	48	1118	0.63
4	5	57	255	841	249	1407	0.59	**4**	1	11	166	1126	171	1475	0.76
5	2	15	149	326	420	912	0.46	**5**	0	0	41	247	628	916	0.68
Total	49	273	1013	1813	849	3997		**Total**	65	338	1041	1681	856	3981	
PA	0.22	0.29	0.44	0.46	0.49			**PA**	0.75	0.66	0.67	0.66	0.73		
KH								**KH**							
Zones	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**	**Zones**	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**
1	52	23	6	20	6	107	0.48	**1**	59	31	3	8	6	107	0.55
2	27	271	98	142	39	577	0.46	**2**	14	421	74	57	11	577	0.72
3	2	94	160	243	68	567	0.28	**3**	0	98	278	172	19	567	0.49
4	0	97	126	628	194	1045	0.6	**4**	0	21	77	817	130	1045	0.78
5	2	44	51	205	303	605	0.5	**5**	0	7	4	114	480	605	0.79
Total	83	529	441	1238	610	2901		**Total**	73	578	436	1168	646	2901	
PA	0.62	0.51	0.36	0.5	0.49			**PA**	0.8	0.72	0.63	0.69	0.74		
HT								**HT**							
Zones	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**	**Zones**	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**
1	50	68	56	54	21	249	0.2	**1**	103	86	24	15	0	228	0.45
2	47	219	195	187	50	698	0.31	**2**	30	378	160	61	7	636	0.59
3	25	145	329	298	84	881	0.37	**3**	2	126	505	236	24	893	0.56
4	13	129	251	490	156	1039	0.47	**4**	1	28	178	736	116	1059	0.69
5	19	51	95	278	281	724	0.38	**5**	0	4	34	178	465	681	0.68
Total	154	612	926	1307	592	3591		**Total**	136	622	901	1226	612	3497	
PA	0.32	0.35	0.35	0.37	0.47			**PA**	0.75	0.6	0.56	0.6	0.75		

**PA – Producer accuracy, UA – User accuracy

Moreover, inclusion of remote sensing data adds more layers (with multiple bands) to the dataset. The additional data is not only spatially variable but also highly relevant to the soil properties as some of the bands (NIR and SWIR) are highly sensitive to soil moisture content, clay minerology and soil organic carbon. Whereas SAR band data obtained from radar sensors provides information about soil texture. Model performance can be thus improved by allowing the model to learn from a diverse set of scenarios with multiple bands correcting for any discrepancies or deficiencies in the initial dataset.

3.2 Regression Model

Similar to the classification model, the regression model also shows an improvement in its performance after including remote sensing data. Regression analysis for actual and predicted yields provided R2 values ranging from 0.28 (HT) to 0.47 (BS) which were increased to 0.67 (HT) to 0.83 (BS) after adding remote sensing data into the model (Table 5Table 5). The scatter plots for all fields showed a noticeable transformation, with data points more densely clustered around the regression line for each field (Figure 4). This visual representation indicates a stronger correlation between predicted and actual yields. Additionally, increase in R^2 values and decrease in RMSE values indicates stronger ability of model to predict yield zones.

Table 5. R-squared and RMSE values of RF regression model for the four fields before (NRS) and after using remote sensing data (RS).

Field	R^2 (NRS)	RMSE (NRS)	R^2 (RS)	RMSE (RS)
BS	0.47	9.60	0.83	5.45
KM	0.3	10.76	0.71	6.7
KH	0.42	3.36	0.71	2.4
HT	0.28	11.86	0.67	7.94

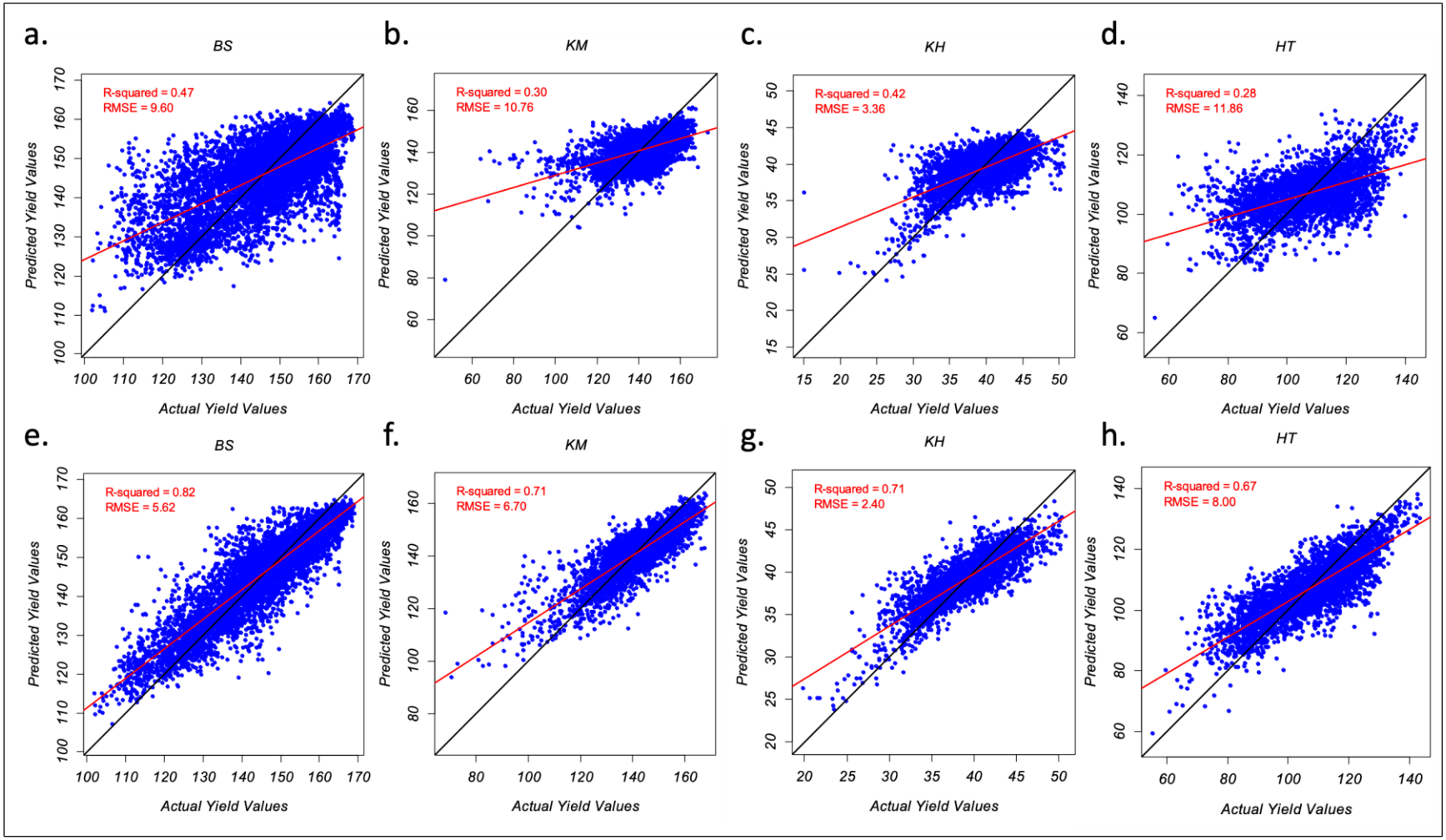

Figure 4. Scatter plots for actual vs predicted yields of four fields using RF regression model before adding remote sensing data (a-d) and after adding remote sensing data (e-h)

RF regression model was used by other studies at different scales ranging from regional level to global level for yield prediction. Prasad et al., (2021) focused on cotton yield prediction utilized long-term agrometeorological and spectral variables, including rainfall, vegetation condition index, standardized precipitation index, growing degree days, and land surface temperature as predictors. This study reported R^2 values ranging from 0.39 to 0.69. In a similar study Jeong, (2016) incorporated climate and biophysical factors such as temperature, evapotranspiration, annual precipitation, soil water content, and soil bulk density, reporting R^2 values between 0.87 and 0.98 while studying crop yield responses across various crops. Another study by Everingham, (2016) focused on sugarcane, using local climate data and biomass indices as features, and reported R^2 values from 0.67 to 0.79, showcasing the model's strong predictive capability. In southeast Australia, an analysis leveraging time series NDVI data along with meteorological variables for yield prediction across three paddocks resulted in R^2 values ranging from 0.45 to 0.87 (Pang et al., 2022). Given the wide variability in R^2 values across multiple studies, the R^2 values obtained in our study ranging from 0.67 to 0.83 are satisfactory for predicting yield values. From these observations it is worth noting that utilization of climate data, biomass data and other environmental data as features potentially increases the model performance. It is evident from the aforementioned studies and the results of our study that the strategic integration of multidimensional data sets not only enriches the model's input but also significantly boosts its performance.

3.3 Variable importance projection (VIP) plots.

VIP plots offers insights into the significance of various features in RF models. In each plot of the classification model, individual features used in the model are represented on the Y- axis and mean decrease gini values on the X- axis (Figure 5). Whereas in regression model, individual features used in the model are represented on Y-axis, percent increase in mean squared error (%IncMSE) and increase in node purity (IncNodePurity) values are represented on X-axis (Figure 6). In both cases, features with higher values on the X-axis in the VIP plots are considered more crucial for the model's predictions.

VIP plots across four fields for RF Classification model demonstrate that the soil properties data [soil hydrology, CEC, WHC, SOM, and texture], consistently exhibit the lowest values for mean decrease gini across all fields (Figure 5). In contrast, the remote sensing features [SWIR, NIR, SAR bands, and NDVI], consistently displayed a higher mean decrease in gini values. This suggests that the soil-related features have limited influence in reducing impurity and achieving class separation within the decision trees of the RF model. Their contribution to accurate classification appears to be relatively minor. The limited impact of soil properties on model performance can be attributed to the lack of variability across the field. When data extracted from corresponding rasters possess uniform pixel values throughout the field, it diminishes the model’s opportunity to learn from these features. Consequently, remote sensing data, like SWIR, NIR, NDVI are consistently high-ranking features. These variables have high mean decrease gini values, which suggests that they are critical in splitting the nodes of the trees within RF model.

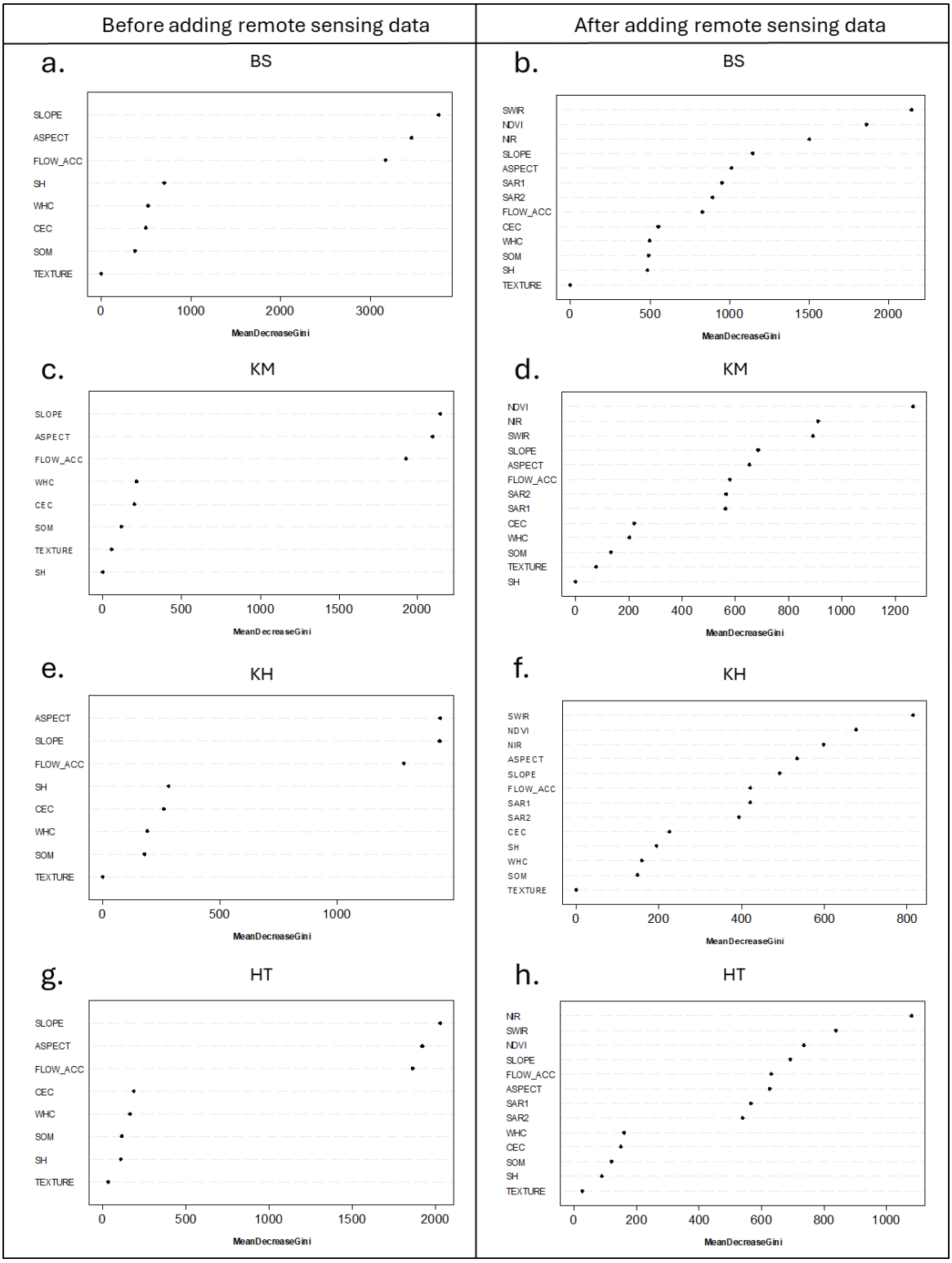

Figure 5. Variable importance projection plots of RF classification model

and thus, have a strong influence on classification performance of the model. Greater variability in remote sensing data aids in reducing impurity and enhancing the accuracy of classification model by accurate categorization of management zones.

In the case of RF regression model VIP plots show a similar pattern as classification model. Features like NDVI, NIR and SWIR being consistently important across the fields, as indicate by their %IncMSE and IncNodePurity values (Figure 6). Given the significance of these features in predicting yield, exclusion of such features deteriorates the model's predictive power. Whereas soil properties like texture have consistently ranked lower among all other features due to their homogeneity within the fields. Features like NDVI, NIR and SWIR have direct relevance to the vegetation status and are strongly correlated to yield. Using features that have strong correlation with the target typically reduces the Mean Square Error (MSE) and often leads to purer nodes in a decision tree. Such features also tend to provide clear and distinct separation of data, enabling the algorithm to make splits that are highly informative for predicting the target leading to more accurate predictions.

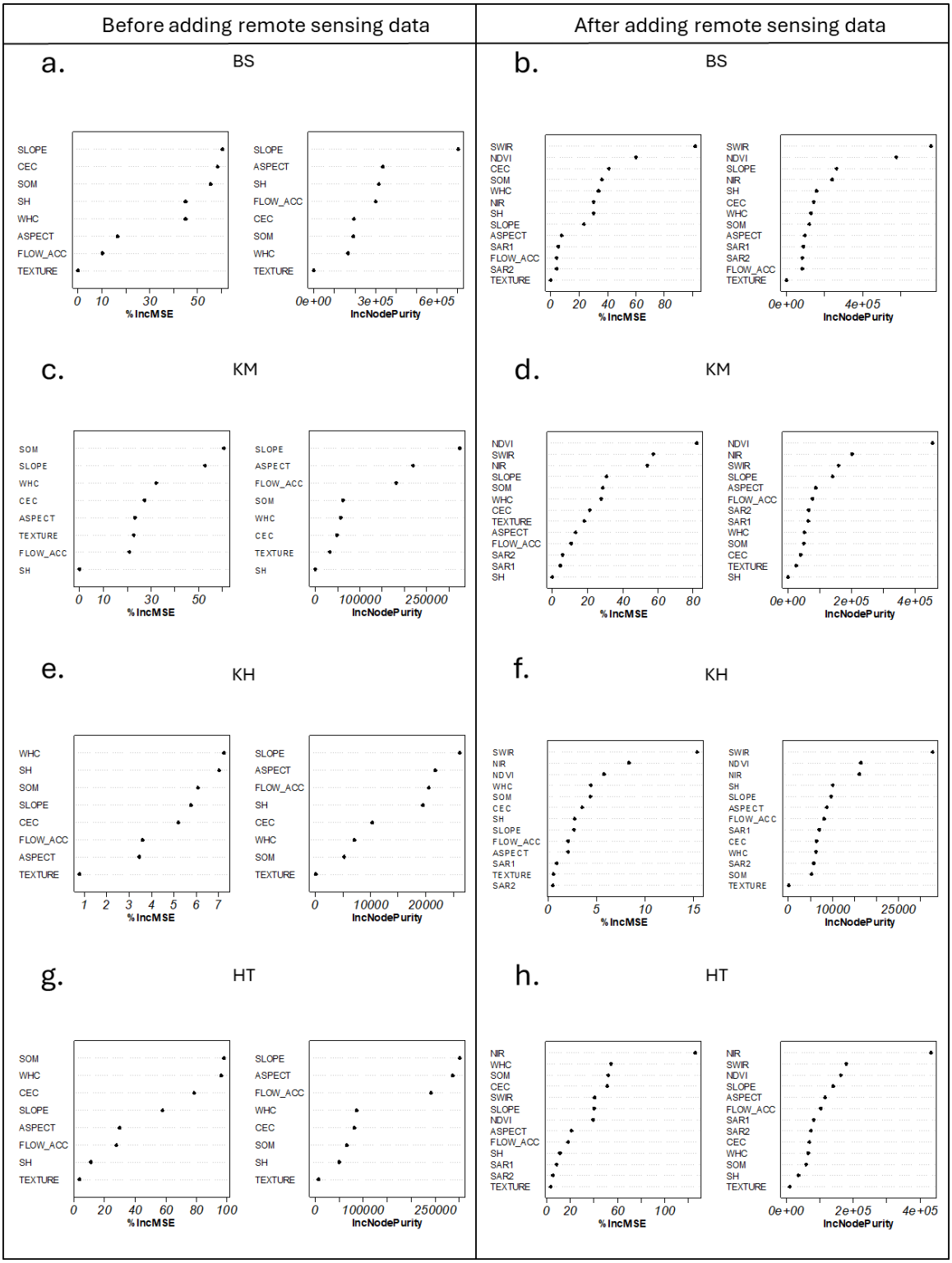

Figure 6. Variable importance projection plots of RF regression model.

3.4 Developing generalized models.

RF classification and regression models mentioned previously were developed at field level for four individual fields (BS, KM, KH, and HT). To develop a generalized model, data from seven corn (BS, HT, KE, KM, KS, LS and MU) and seven soybean fields (AD, HT, LH, LN, LW, KH and MU) were selected to further analyze and develop more generalized classification and regression models. The combined data was then used for training and testing the models separately by crop type to ensure reliability in predicting management zones. Results from these models are discussed separately below.

3.4.1 Classification model

After combining the data for seven fields separately for corn and soybean, OA are 0.75 and 0.8 respectively reflecting a higher accuracy for the combined data set in predicting management zones. Kappa scores are 0.59 for corn and 0.71 for soybean suggesting a substantial agreement level. F-1 scores are 0.93 and 0.95 respectively for corn and soybean indicating a good balance between precision and recall in the classification process (Table 6).

Table 6. Confusion matrices for Corn and Soybean combined data

Soybean								**Corn**							
Zones	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**	**Zones**	**1**	**2**	**3**	**4**	**5**	**Total**	**UA**
1	217	124	13	10	3	367	0.59	**1**	260	74	7	39	9	389	0.66
2	33	1207	291	58	21	1610	0.74	**2**	36	1054	162	224	37	1513	0.69
3	3	240	1308	363	99	2013	0.64	**3**	5	242	521	598	85	1451	0.35
4	1	25	230	1406	486	2148	0.65	**4**	5	118	139	6433	836	7531	0.85
5	0	1	24	217	5295	5537	0.95	**5**	1	10	6	980	2420	3417	0.7
Total	254	1597	1866	2054	5904	11675		**Total**	307	1498	835	8247	3387	14301	
PA	0.85	0.75	0.7	0.68	0.89			**PA**	0.84	0.7	0.62	0.77	0.71		
OA							0.8	**OA**							0.75
KA							0.71	**KA**							0.59
F-1							0.93	**F-1**							0.95

OA- Overall accuracy; KA- Kappa score; F-1 – F-1 score; PA -Producer accuracy; UA- User accuracy.

3.4.2 Regression model

The regression model for the combined data set had R^2 values of 0.8 for corn and 0.85 for soybean indicating a strong positive linear relationship between actual and predicted yield values (Figure 7). RMSE measures the discrepancy between the values predicted by the model and the actual values observed. Overall regression model for corn and soybean had high R^2 values and lower RMSE values indicating a significant reliability in the predictions.

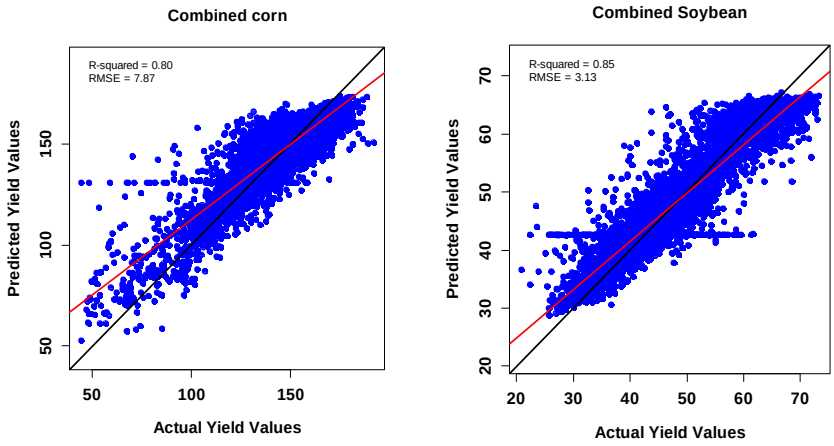

Figure 7. Scatter plots of actual vs predicted yields for combined corn (left) and soybean (right).

A noticeable increase has been observed in the performance of both classification (OA, KA, and F-1) and regression model (R^2, RMSE) after integrating data across seven fields. Aggregating data from multiple fields introduces a more diverse set of soil conditions and plant responses, presenting the models with a broader spectrum of scenarios to learn from. The data points for individual fields range between 9,669- 23,342 points whereas after combining the seven field's data together, total data points obtained for corn and soybean were 47,947 and 38,918 respectively. Increase in data points contributes to a more robust learning process, where the models are less likely to overfit

to the peculiarities of a single field and are more capable of capturing the underlying patterns common across different field conditions. Moreover, developing separate models for corn and soybean while combining data from multiple fields allows crop specific calibration of features. Since each crop has unique responses to the different features used in the model, allowing effective model training. Dataset resulting from the amalgamation of multiple fields provides a comprehensive overview, ensuring that the models developed are not only trained on a larger volume of data but also encapsulate a wide array of field conditions. Such models are inherently more adaptable at providing reliable predictions across various locations and conditions.

When comparing the results of both classification and regression model corn data got lower OA, KA, and F-1 score in case of classification and lower R^2, higher RMSE values compared to soybean data. The inherent difference in the yield ranges across the seven corn fields led to inaccurate classification of management zones where the yield values within specific management zones are different across the fields. However, soybean yield ranges were similar across the seven fields leading to more accurate classification of management zones.

3.5 Visual comparison of all predicted maps

Prediction maps were generated for visual comparison of the model performance at each stage of model development. Figure 8 illustrates the actual map of KM field (multiyear yield average map obtained from SMS Ag software) alongside the prediction maps produced by both classification and regression models before and after incorporating remote sensing data along with the prediction maps generated based on the combined corn data. Prediction map generated by classification and regression models

before incorporating remote sensing data lacks distinct zone boundaries and misclassifies a majority of data points of zones 3 and 4 as zone 5. Conversely, prediction map generated by classification and regression models after incorporating remote sensing data showcases a significant improvement and more accurate classification of zones, offering a clear and more acceptable delineation of different management zones. A similar pattern was observed across all the fields.

Although OA and R^2 of combined corn data were higher for both classification and regression models compared to the models for individual fields, misclassification of multiple management zones was profound in the prediction maps generated. However, among the two models, the management zones in the prediction map generated by regression model were closer to the actual map in case of KM field (Figure 8).

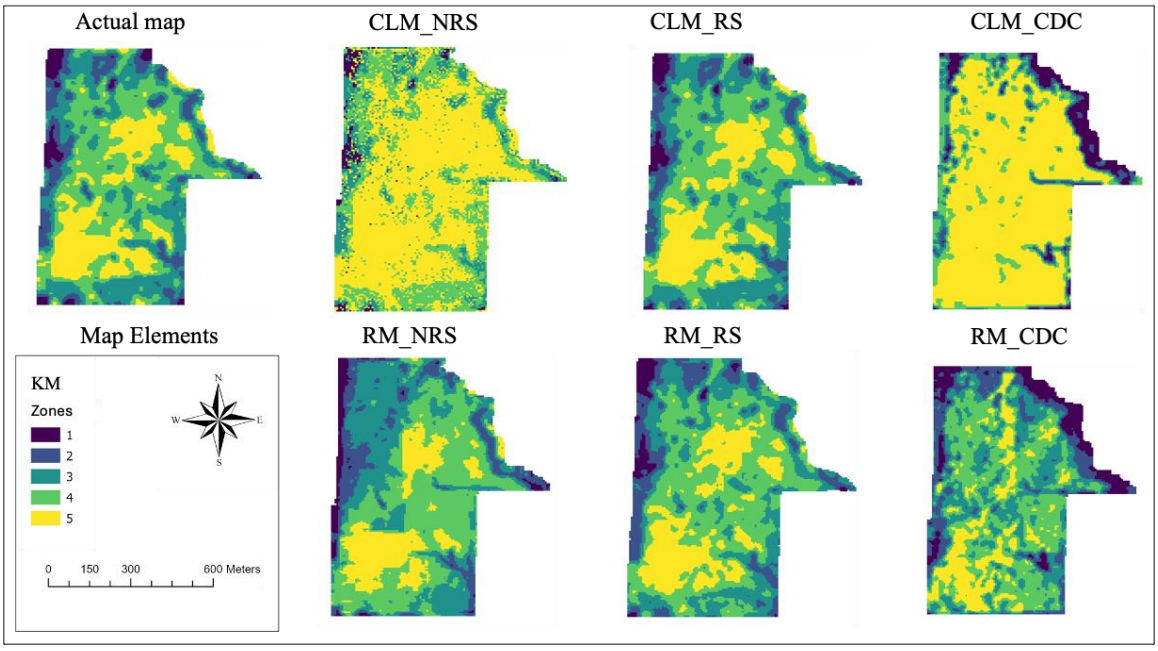

Figure 8. Actual and prediction maps for KM field.

CLM- Classification model
NRS – Before adding remote sensing data
CDC- Combined data for Corn
RM – Regression model
RS – After adding remote sensing data.

While comparing prediction maps of various fields for both corn [BS (Figure 9), KS (Figure 18), LS, MU (Figure 19)] and soybean [KH (Figure 11), AD, HI (Figure 13), LH, LN (Figure 15), LW, MU (Figure 16)], it becomes evident that the prediction maps generated by regression models exhibit a closer alignment with the actual maps compared to classification models. This observation underscores the effectiveness of regression models in capturing the spatial variability of yield. From these results it becomes evident that continuous prediction of yield values by regression provides better results compared to categorical prediction of management zones by classification.

3.6 Model deployment

After training and testing the classification and regression models developed from combined corn and soybean data, these models were further evaluated on two new fields for corn (FN and AT) and soybean (CW and PO). Upon observing the prediction maps (Figure 17 ,Figure 14) generated by these models it can be interpreted that both models fail to classify the management zones accurately. This stems from various factors, including limitations in model transferability due to differing environmental conditions and site-specific factors.

Addition of high-resolution data with more spatial variability, features that are more correlated to yield such as environmental factors, biomass indices would enhance the predictive power of models used in agricultural management. These additional features offer a more comprehensive representation of the factors influencing crop yield, allowing models to capture subtle variations that may significantly impact yield. By integration of environmental variables such as climate data, models gain insights into the complex interactions between the environment and crop growth by enabling models to

account for spatial heterogeneity under different growing conditions, thus improving their ability to accurately predict yield variability across different regions and fields.

Incorporation of high-resolution data, and correlated features not only enriches the input dataset but also facilitates model's ability to learn complex relationships between predictor variables and yield outcomes. By training on diverse datasets that encompass a wide range of environmental conditions models become more robust and adaptable, enhancing their predictive power and generalization capabilities.

CHAPTER 4: CONCLUSION

This study advances the development of comprehensive machine learning models that integrate a wide array of data encompassing landscape data, soil properties data and remote sensing data for optimal selection of soil sampling locations. This was achieved by creating management zones using random forest regression and classification of models. The results indicated a substantial improvement in model performance with the incorporation of remote sensing data, that can be seen in increased overall accuracy on an average of 24%, kappa scores on an average of 33%, and F-1 scores on an average of 1 to 4% for the classification model across four fields. In case of regression, R^2 values increased from range of 0.28 - 0.47 to 0.67 to 0.83. These enhancements highlight the significant role of remote sensing data in capturing spatial variability of soil properties and yields. Incorporating remote sensing data including NIR, SWIR, NDVI and SAR bands add more layers to the data set and this additional data which is spatially variable across the field allowing the model to have a more comprehensive set of scenarios. This broader exposure helps to fine tune model's decision boundaries to handle complex, nonlinear relationships more effectively, leading to more accurate predictions.

Furthermore, increase in accuracy levels and R^2 values of classification and regression analysis of a generalized model developed by combining data from seven corn fields and seven soybeans showcased the model's robustness and adaptability. Upon observing prediction maps generated by these models it is worth noting that despite having higher accuracies compared to individual fields, the generalized model fails to delineate the management zones accurately. This might to due to limitations of model's

transferability, different environmental conditions, and site-specific features across different fields or a lack of predictor covariates.

Overall comparison of prediction maps generated by the models in this study reveals that the management zones generated by regression models are closely aligned to actual maps. This underscores that prediction of continuous yield values by regression provides better results than categorical prediction of management zones by using classification model.

This study acknowledges limitations in model transferability to new fields, suggesting the need for incorporating more diverse and spatially variable data to enhance predictive power. A diverse dataset provides a comprehensive overview and provides more scope for models to learn and capture the underlying patterns, complex relationship between the features across the different fields. Moreover, addition of features that are more correlated to target such as environmental data, biomass indices, high resolution soil properties data and considering the fields that have more landscape variations aids the model training on wide array of field conditions which increase the model's adaptability and provides reliable predictions across various location and conditions.

Future works includes development of more robust models considering wide array of data sets including environmental variables, biomass indices, remote sensing data, high-resolution soil properties and landscape data for accurate prediction of management zones. These management zones can then be utilized to generate sampling points across the fields. As these zones are created by considering spatial heterogeneity of soil properties across the landscape, the samples collected from across these zones would represent the entire field. By providing information about the field conditions in the form

of management zones it can aid the producers to improve their resource use efficiency by focusing sampling efforts where they are most needed. It also provides various options to the producers regarding quantity of samples to be collected from each zone. This study has the potential to transform the way sampling is carried out with seamless integration of humans, technology, and data. The outcomes from this study will increase the precision and efficacy of sampling, which will reduce efforts, expenses, and improves the reliability of soil test results.

LITERATURE CITED

Abdulraheem, M. I., Zhang, W., Li, S., Moshayedi, A. J., Farooque, A. A., & Hu, J. (2023). Advancement of Remote Sensing for Soil Measurements and Applications: A Comprehensive Review. *Sustainability*, *15*(21). https://doi.org/10.3390/su152115444

Ahmad, S., Kalra, A., & Stephen, H. (2009). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources*, *33*(2010), 69. https://doi.org/10.1016/j.advwatres.2009.10.008

Ahmadi, A., Emami, M., Daccache, A., & He, L. (2021). Soil Properties Prediction for Precision Agriculture Using Visible and Near-Infrared Spectroscopy: A Systematic Review and Meta-Analysis. *Agronomy*, *11*(3), 433. https://doi.org/10.3390/agronomy11030433

An, Y., Yang, L., Zhu, A. X., Qin, C., & Shi, J. J. (2018). Identification of representative samples from existing samples for digital soil mapping. *Geoderma*, *311*, 109–119. https://doi.org/10.1016/j.geoderma.2017.03.014

Asare, E., & Segarra, E. (2018). Adoption and extent of adoption of georeferenced grid soil sampling technology by cotton producers in the southern US. *Precision Agriculture*, *19*(6), 992–1010. https://doi.org/10.1007/S11119-018-9568-3

Auernhammer, H. (2001). Precision farming — the environmental challenge. *Computers and Electronics in Agriculture*, *30*(1–3), 31–43. https://doi.org/10.1016/S0168-1699(00)00153-8

Ayele, G. T., Demissie, S. S., Tilahun, S. A., Jeong, J., & Jemberie, M. A. (2015). Assessing drought severity from Multi-Temporal GIMMSNDVI and rainfall interactions. *36th Hydrology and Water Resources Symposium: The Art and Science of Water, Engineers Australia* , 306–314. http://hdl.handle.net/10072/124146

Balaram, V., & Sawant, S. S. (2022). Indicator Minerals, Pathfinder Elements, and Portable Analytical Instruments in Mineral Exploration Studies. *Minerals*, *12*(4), 394. https://doi.org/10.3390/min12040394

Blackmore, S. (2000). The interpretation of trends from multiple yield maps. *Computers and Electronics in Agriculture*, *26*(1), 37–51. https://doi.org/10.1016/S0168-1699(99)00075-7

Bondi, G., Creamer, R., Ferrari, A., Fenton, O., & Wall, D. (2018). Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. *Geoderma*, *318*, 137–147. https://doi.org/10.1016/j.geoderma.2017.11.035

Boswell, M. T., & Patil, G. P. (1987). A perspective of composite sampling. *Communications in Statistics - Theory and Methods*, *16*(10), 3069–3093. https://doi.org/10.1080/03610928708829558

Brady, N. C., & Weil, R. R. (2016). Nature and Properties of Soils. In *Pearson Education* (15th ed.). Pearson Education.

Brady, N., & Weil, R. (2004). *Elements of the nature and properties of soils*. http://faculty.washington.edu/zabow/ESC210/210Syllabus_2007.doc

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brevik, E. C., Cerdà, A., Mataix-Solera, J., Pereg, L., Quinton, J. N., Six, J., & Van Oost, K. (2015). The interdisciplinary nature of SOIL. *Soil*, *1*(1), 117–129. https://doi.org/10.5194/SOIL-1-117-2015

Brevik, E. C., Homburg, J. A., Miller, B. A., Fenton, T. E., Doolittle, J. A., & Indorante, S. J. (2016). Selected highlights in American soil science history from the 1980s to the mid-2010s. *Catena*, *146*, 128–146. https://doi.org/10.1016/j.catena.2016.06.021

Carter, M. R., & Gregorich, E. G. (2007). Soil Sampling and Methods of Analysis. In *Soil Sampling and Methods of Analysis: Second Edition* (2nd ed.). CRC Press.

Ceddia, M. B., Vieira, S. R., Villela, A. L. O., Mota, L. D. S., Anjos, L. H. C. Dos, & Carvalho, D. F. De. (2009). Topography and spatial variability of soil physical properties. *Scientia Agricola*, *66*(3), 338–352. https://doi.org/10.1590/S0103-90162009000300009

Chang, J., Clay, D. E., Carlson, C. G., Clay, S. A., Malo, D. D., Berg, R., Kleinjan, J., & Wiebold, W. (2003). Different Techniques to Identify Management Zones Impact Nitrogen and Phosphorus Sampling Variability. *Agronomy Journal*, *95*(6), 1550–1559. https://doi.org/10.2134/agronj2003.1550

Chen, L. F., He, Z. Bin, Du, J., Yang, J. J., & Zhu, X. (2016). Patterns and environmental controls of soil organic carbon and total nitrogen in alpine ecosystems of northwestern China. *Catena*, *137*, 37–43. https://doi.org/10.1016/j.catena.2015.08.017

Clay, D. E., Chang, J., & Gregg Carlson, C. (2019). *Precision Soil Sampling*. IGrow Corn: Best Management Practices for Corn Production. https://extension.sdstate.edu/igrow-corn-best-management-practices-corn-production

Cline, M. G. (1944). Principles of Soil Sampling. *Soil Science*, *58*(4), 275–288. https://doi.org/10.1097/00010694-194410000-00003

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). John Wiley & Sons.

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture*, *46*(1–3), 11–43. https://doi.org/10.1016/j.compag.2004.10.005

Diaz, G. I., Fokoue, N. A., Nannicini, G., & Samulowitz, H. (2017). An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development*, *61*(4). https://doi.org/10.1147/jrd.2017.2709578

Diker, K., Heermann, D. F., & Brodahl, M. K. (2004). Frequency analysis of yield for delineating yield response zones. *Precision Agriculture*, *5*(5), 435–444. https://doi.org/10.1007/S11119-004-5318-9/metrics

Dinkins, Courtney. Pariera., Jones, Clain., & Olson-Rutz, K. (2008). Soil Sampling Strategies. *A Self-Learning Resource from MSU Extension*. https://doi.org/MT200803AG New, 4(08)

Domenech, M. B., Amiotti, N. M., Costa, J. L., & Castro-Franco, M. (2020). Prediction of topsoil properties at field-scale by using C-band SAR data. *International Journal of Applied Earth Observation and Geoinformation*, *93*, 102–197. https://doi.org/10.1016/j.jag.2020.102197

Doran, J. W. (2002). Soil health and global sustainability: translating science into practice. *Agriculture, Ecosystems & Environment*, *88*(2), 119–127. https://doi.org/10.1016/S0167-8809(01)00246-8

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology*, 3–11. https://doi.org/10.1007/978-3-319-18305-3_1

Everingham, Y., Sexton, J., Skocaj, D., & Inman Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, *36*(2), 1–9. https://doi.org/10.1007/S13593-016-0364

Fathololoumi, S., Vaezi, A. R., Alavipanah, S. K., Ghorbani, A., Saurette, D., & Biswas, A. (2021). Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach. *Geoderma*, *385*. https://doi.org/10.1016/j.geoderma.2020.114901

Flint, A. L., & Flint, L. E. (2002). 2.2 Particle Density. In J. H. Dane, G. C. Topp, & G. S. Campbell (Eds.), *Methods of Soil Analysis: Part 4 Physical methods* . Soil Science Society of America.

Flowers, M., Weisz, R., & White, J. G. (2005). Yield-Based Management Zones and Grid Sampling Strategies: Describing Soil Test and Nutrient Variability. *Agronomy Journal*, *97*(3), 968–982. https://doi.org/10.2134/agronj2004.0224

Franzen, D., Halvorson, A., & Hofman, V. (2000). Management zones for soil N and P levels in the Northern Great Plains. *Management Zones for Soil N and P Levels in the Northern Great Plains.*, 1–10. https://www.cabdirect.org/cabdirect/abstract/20023117465

Franzen, D. W., & Peck, T. R. (1995). Field Soil Sampling Density for Variable Rate Fertilization. *Journal of Production Agriculture*, *8*(4), 568–574. https://doi.org/10.2134/jpa1995.0568

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining* (1st ed., Vol. 72). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-10247-4

Gili, A., Álvarez, C., Bagnato, R., & Noellemeyer, E. (2017). Comparison of three methods for delineating management zones for site-specific crop management. *Computers and Electronics in Agriculture*, *139*, 213–223. https://doi.org/10.1016/J.compag.2017.05.022

González, C., Mira-McWilliams, J., & Juárez, I. (2015). Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission & Distribution*, *9*(11), 1120–1128. https://doi.org/10.1049/iet-gdt.2014.0655

Greenwell, B. M., Boehmke, B. C., & Gray, B. (2020). Variable Importance Plots-An Introduction to the vip Package. *The R Journal*, *13*(1), 343. https://journal.r-project.org/articles/RJ-2020-013/RJ-2020-013.pdf

Han, H., Guo, X., & Yu, H. (2016). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *0*, 219–224. https://doi.org/10.1109/icsess.2016.7883053

Havlin, J. L. (2020). Soil: Fertility and Nutrient Management. *Landscape and Land Capacity*, 251–265. https://doi.org/10.1201/9780429445552-34

Hemingway, R. G. (1955). Soil-sampling errors and advisory analyses. *The Journal of Agricultural Science*, *46*(1), 1–8. https://doi.org/10.1017/S0021859600039563

Hengl, T., Rossiter, D. G., & Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Research*, *41*(8), 1403–1422. https://doi.org/10.1071/sr03005

Jakšík, O., Kodešová, R., Kubiš, A., Stehlíková, I., Drábek, O., & Kapička, A. (2015). Soil aggregate stability within morphologically diverse areas. *Catena*, *127*, 287–299. https://doi.org/10.1016/j.catena.2015.01.010

Jenny, H. (1994). *Factors of soil formation : a system of quantitative pedology*. Dover Publications, Inc.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *Plos One*, *11*(6). https://doi.org/10.1371/journal.pone.0156571

Kazemi, M., & Samavati, F. F. (2023). Automatic Soil Sampling Site Selection in Management Zones Using a Multi-Objective Optimization Algorithm. *Agriculture (Switzerland)*, *13*(10). https://doi.org/10.3390/agriculture13101993

Kerry, R., Oliver, M. A., & Frogbrook, Z. L. (2010). Sampling in Precision Agriculture. *Geostatistical Applications for Precision Agriculture*, 35–63. https://doi.org/10.1007/978-90-481-9133-8_2

Khosla, R., Inman, D., Westfall, D. G., Reich, R. M., Frasier, M., Mzuku, M., Koch, B., & Hornung, A. (2008). A synthesis of multi-disciplinary research in precision agriculture: Site-specific management zones in the semi-arid western Great Plains of the USA. *Precision Agriculture*, *9*(1–2), 85–100. https://doi.org/10.1007/S11119-008-9057-1

Knowles, O., & Dawson, A. (2018). CURRENT SOIL SAMPLING METHODS-A REVIEW. *Farm Environmental Planning - Science, Policy and Pracice*. http://flrc.massey.ac.nz/publications.html.

Kumhálová, J., Kumhála, F., Kroulík, M., & Matějková, Š. (2011). The impact of topography on soil properties and yield and the effects of weather conditions. *Precision Agriculture*, *12*(6), 813–830. https://doi.org/10.1007/S11119-011-9221-X

Lal, R. (2008). Soils and sustainable agriculture. A review. *Agronomy for Sustainable Development*, *28*(1), 57–64. https://doi.org/10.1051/AGRO:2007025/METRICS

Lark, R. M., & Stafford, J. V. (1997). Classification as a first step in the interpretation of temporal and spatial variation of crop yield. *Annals of Applied Biology*, *130*(1), 111–121. https://doi.org/10.1111/J.1744-7348.1997.TB05787.X

Lark, R. M., & Stafford, J. V. (1998). Information on within-field variability from sequences of yield maps: Multivariate classification as a first step of interpretation. *Nutrient Cycling in Agroecosystems*, *50*(1–3), 277–281. https://doi.org/10.1007/978-94-017-3021-1_27/COVER

Lawrence, P. G., Roper, W., Morris, T. F., & Guillard, K. (2020). Guiding soil sampling strategies using classical and spatial statistics: A review. *Agronomy Journal*, *112*(1), 493–510. https://doi.org/10.1002/AGJ2.20048

Lybrand, R. A., & Rasmussen, C. (2015). Quantifying Climate and Landscape Position Controls on Soil Development in Semiarid Ecosystems. *Soil Science Society of America Journal*, *79*(1), 104–116. https://doi.org/10.2136/SSSAJ2014.06.0242

Madow, W. G., & Madow, L. H. (1944). On the Theory of Systematic Sampling, I. *The Annals of Mathematical Statistics*, *15*(1), 1–24. https://doi.org/10.1214/AOMS/1177731312

Mallarino, A. P. (2005). Testing of Soils. *Encyclopedia of Soils in the Environment*, *4*, 143–149. https://doi.org/10.1016/B0-12-348530-4/00302-7

Mallarino, A. P. (2023). Testing for nutrients in the soil and the environment. *Reference Module in Earth Systems and Environmental Sciences*. https://doi.org/10.1016/B978-0-12-822974-3.00197-X

Mallarino, A. P., & Wittry, D. J. (2004). Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter. *Precision Agriculture*, *5*(2), 131–144. https://doi.org/10.1023/B:PRAG.0000022358.24102.1B/METRICS

Mayer, P., Wania, F., & Wong, C. S. (2014). Advancing passive sampling of contaminants in environmental science. *Environmental Science: Processes & Impacts*, *16*(3), 366–368. https://doi.org/10.1039/C4EM90004A

McBratney, A. B., Santos, M. L. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney, A., Gruijter, Jaap. D., & Bryce, A. (2019). Pedometrics timeline. *Geoderma*, *338*, 568–575. https://doi.org/10.1016/j.geoderma.2018.11.048

Meisinger, J. J. (2015). Evaluating Plant-Available Nitrogen in Soil-Crop Systems. *Nitrogen in Crop Production*, 389–416. https://doi.org/10.2134/1990.nitrogenincropproduction.C26

Meyer, W. B., & Turner, B. L. (2003). Human Population Growth and Global Land-Use/Cover Change. *Annual Review of Ecology, Evolution, and Systematics*, *23*(1), 39–61. https://doi.org/10.1146/annurev.es.23.110192.000351

Miao, Y., Mulla, D. J., & Robert, P. C. (2018). An integrated approach to site-specific management zone delineation. *Frontiers of Agricultural Science and Engineering*, *5*(4), 432–441. https://doi.org/10.15302/j-fase-2018230

Miller, B. A., & Schaetzl, R. J. (2015). Digital Classification of Hillslope Position. *Soil Science Society of America Journal*, *79*(1), 132–145. https://doi.org/10.2136/sssaj2014.07.0287

Milne, A., Webster, R., Ginsburg, D., in, D. K.-C. and electronics, & 2012, undefined. (n.d.). Spatial multivariate classification of an arable field into compact management zones based on past crop yields. *Elsevier*. Retrieved March 3, 2024, from https://www.sciencedirect.com/science/article/pii/S0168169911002353

Minasny, B., McBratney, A. B., & Walvoort, D. J. J. (2007). The variance quadtree algorithm: Use for spatial sampling design. *Computers & Geosciences*, *33*(3), 383–392. https://doi.org/10.1016/j.cageo.2006.08.009

Motia, S., & Reddy, S. R. N. (2021). Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. *Journal of Physics: Conference Series*, *1950*(1), 012037. https://doi.org/10.1088/1742-6596/1950/1/012037

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., & Mouazen, A. M. (2017). Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review. *Advances in Agronomy*, *143*, 175–245. https://doi.org/10.1016/BS.AGRON.2017.01.003

Nawar, S., & Mouazen, A. M. (2018). Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. *Computers and Electronics in Agriculture*, *151*, 469–477. https://doi.org/10.1016/J.COMPAG.2018.06.042

Obi, J. C., Ogban, P. I., Ituen, U. J., & Udoh, B. T. (2014). Development of pedotransfer functions for coastal plain soils using terrain attributes. *Catena*, *123*, 252–262. https://doi.org/10.1016/J.CATENA.2014.08.015

Oliver, Y. M., Robertson, M. J., & Wong, M. T. F. (2010). Integrating farmer knowledge, precision agriculture tools, and crop simulation modelling to evaluate management options for poor-performing patches in cropping fields. *European Journal of Agronomy*, *32*(1), 40–50. https://doi.org/10.1016/J.EJA.2009.05.002

Padarian, J., Minasny, B., & McBratney, A. B. (2020a). Machine learning and soil sciences: A review aided by machine learning tools. *Soil*, *6*(1), 35–52. https://doi.org/10.5194/SOIL-6-35-2020

Padarian, J., Minasny, B., & McBratney, A. B. (2020b). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, *6*(1), 35–52. https://doi.org/10.5194/SOIL-6-35-2020

Pang, A., Chang, M. W. L., & Chen, Y. (2022). Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia. *Sensors 2022, Vol. 22, Page 717*, *22*(3), 717. https://doi.org/10.3390/S22030717

Park, S. J., & Vlek, P. L. G. (2002). Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. *Geoderma*, *109*(1–2), 117–140. https://doi.org/10.1016/S0016-7061(02)00146-5

Peck, T. R. (1990). Soil testing: Past, present and future. *Communications in Soil Science and Plant Analysis*, *21*(13–16), 1165–1186. https://doi.org/10.1080/00103629009368297

Pennock, D., Yates, T., & Braidek, J. (2008). Soil sampling designs. In *Soil sampling and methods of analysis* (Vol. 2, pp. 25–37). https://www.niordc.ir/uploads/86_106_Binder1.pdf#page=25

Penuelas, Josep., Coello, Fernando., & Sardans, Jordi. (2023). A better use of fertilizers is needed for global food security and environmental sustainability. *Agriculture and Food Security*, *12*(1), 1–9. https://doi.org/10.1186/S40066-023-00409-5

Petropoulos, G. P., Ireland, G., & Barrett, B. (2015). Surface soil moisture retrievals from remote sensing: Current status, products & future trends. *Physics and Chemistry of the Earth, Parts A/B/C*, *83–84*, 36–56. https://doi.org/10.1016/J.PCE.2015.02.009

Petrovskaia, A., Ryzhakov, G., & Oseledets, I. (2021). Optimal soil sampling design based on the maxvol algorithm. *Geoderma*, *402*, 115362. https://doi.org/10.1016/j.geoderma.2021.115362

Pham, T. H., Acharya, P., Bachina, S., Osterloh, K., & Nguyen, K. D. (2024). Deep-learning framework for optimal selection of soil sampling sites. *Computers and Electronics in Agriculture*, *217*. https://doi.org/10.1016/j.compag.2024.108650

Pierson, F. B., & Mulla, D. J. (1990). Aggregate Stability in the Palouse Region of Washington: Effect of Landscape Position. *Soil Science Society of America Journal*, *54*(5), 1407–1412. https://doi.org/10.2136/SSSAJ1990.03615995005400050033X

Posit team. (2023). *RStudio: Integrated Development Environment for R* (2023.0.0.463).

Prasad, N. R., Patel, N. R., & Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*, *29*(2), 195–206. https://doi.org/10.1007/S41324-020-00346-6/METRICS

Reeves, J. L., & Liebig, M. A. (2016). Depth Matters: Soil pH and Dilution Effects in the Northern Great Plains. *Soil Science Society of America Journal*, *80*(5), 1424–1427. https://doi.org/10.2136/SSSAJ2016.02.0036N

Reza, S. K., Nayak, D. C., Chattopadhyay, T., Mukhopadhyay, S., Singh, S. K., & Srinivasan, R. (2015). Spatial distribution of soil physical properties of alluvial soils: a geostatistical approach. *Archives of Agronomy and Soil Science*, *62*(7), 972–981. https://doi.org/10.1080/03650340.2015.1107678

Savci, S. (2012). Investigation of Effect of Chemical Fertilizers on Environment. *APCBEE Procedia*, *1*, 287–292. https://doi.org/10.1016/j.apcbee.2012.03.047

Seibert, J., Stendahl, J., & Sørensen, R. (2007). Topographical influences on soil properties in boreal forests. *Geoderma*, *141*(1–2), 139–148. https://doi.org/10.1016/J.GEODERMA.2007.05.013

Sengupta, A., & Banerjee, H. (2012). Soil-less culture in modern agriculture. *World Journal of Science and Technology*, *2012*(7), 103–108.

Shaner, D. L., Khosla, R., Brodahl, M. K., Buchleiter, G. W., & Farahani, H. J. (2008). How Well Does Zone Sampling Based on Soil Electrical Conductivity Maps Represent Soil Variability? *Agronomy Journal*, *100*(5), 1472–1480. https://doi.org/10.2134/AGRONJ2008.0060

Soil Survey Geographic Database (SSURGO) | Natural Resources Conservation Service. (2022). https://www.nrcs.usda.gov/resources/data-and-reports/soil-survey-geographic-database-ssurgo

Sparrow, L. A., Peverill, Kenneth. Ian., & Reuter, Douglas. J. (1999). *Soil Analysis: An Interpretation Manual*. CSIRO Pub.

Tan, Kim. Howard. (2005). *Soil sampling, preparation, and analysis* (2nd ed.). CRC Press. https://doi.org/https://doi.org/10.1201/9781482274769

Tisdale, S., Nelson, W., & Beaton, J. (1985). *Soil fertility and fertilizers.* https://www.cabdirect.org/cabdirect/abstract/19851998321

Tomaz, A., Martins, I., Catarino, A., Mourinha, C., Dôres, J., Fabião, M., Boteta, L., Coutinho, J., Patanita, M., & Palma, P. (2022). Insights into the Spatial and Temporal Variability of Soil Attributes in Irrigated Farm Fields and Correlations with Management Practices: A Multivariate Statistical Approach. *Water 2022, Vol. 14, Page 3216*, *14*(20), 3216. https://doi.org/10.3390/W14203216

Walker, P. (1968). *Hill-slope models and soil formation. I. Open systems.* https://www.cabidigitallibrary.org/doi/full/10.5555/19701904671

Wang, J. F., Stein, A., Gao, B. B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*(1), 1–14. https://doi.org/10.1016/J.SPASTA.2012.08.001

Webster, R. (2000). Is soil variation random? *Geoderma*, *97*(3–4), 149–163. https://doi.org/10.1016/S0016-7061(00)00036-7

Weerts, H. J. P., Mueller, A. C., & Vanschoren, J. (2020). *Importance of Tuning Hyperparameters of Machine Learning Algorithms*. https://arxiv.org/abs/2007.07588v1

Wolf, B. (1999). *The fertile triangle :the interrelationship of air, water, and nutrients in maximizing soil productivity*. Food Products Press.

Yang, H., Yoo, H., Lim, H., Kim, J., & Choi, H. T. (2021). Impacts of soil properties, topography, and environmental features on soil water holding capacities (SWHCs) and their interrelationships. *Land*, *10*(12), 1290. https://doi.org/10.3390/land10121290/S1

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zhu, M., Feng, Q., Qin, Y., Cao, J., Zhang, M., Liu, W., Deo, R. C., Zhang, C., Li, R., & Li, B. (2019). The role of topography in shaping the spatial patterns of soil organic carbon. *Catena*, *176*, 296–305. https://doi.org/10.1016/j.catena.2019.01.029

APPENDIX I : PREDICTION MAPS

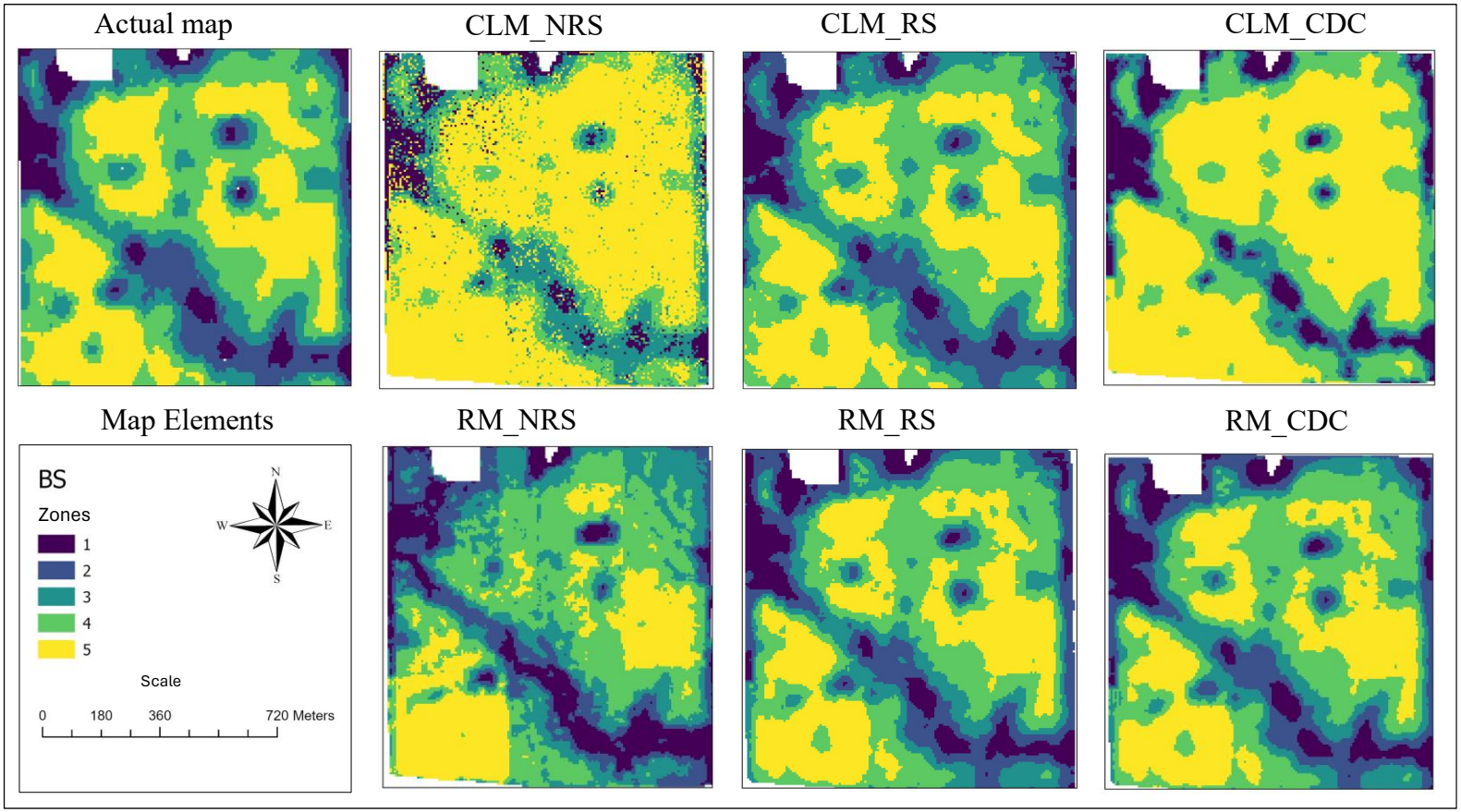

Figure 9. Actual and prediction maps for BS field.
CLM- Classification model RM – Regression model CDC- Combined data for Corn
NRS – Before adding remote sensing data RS – After adding remote sensing data

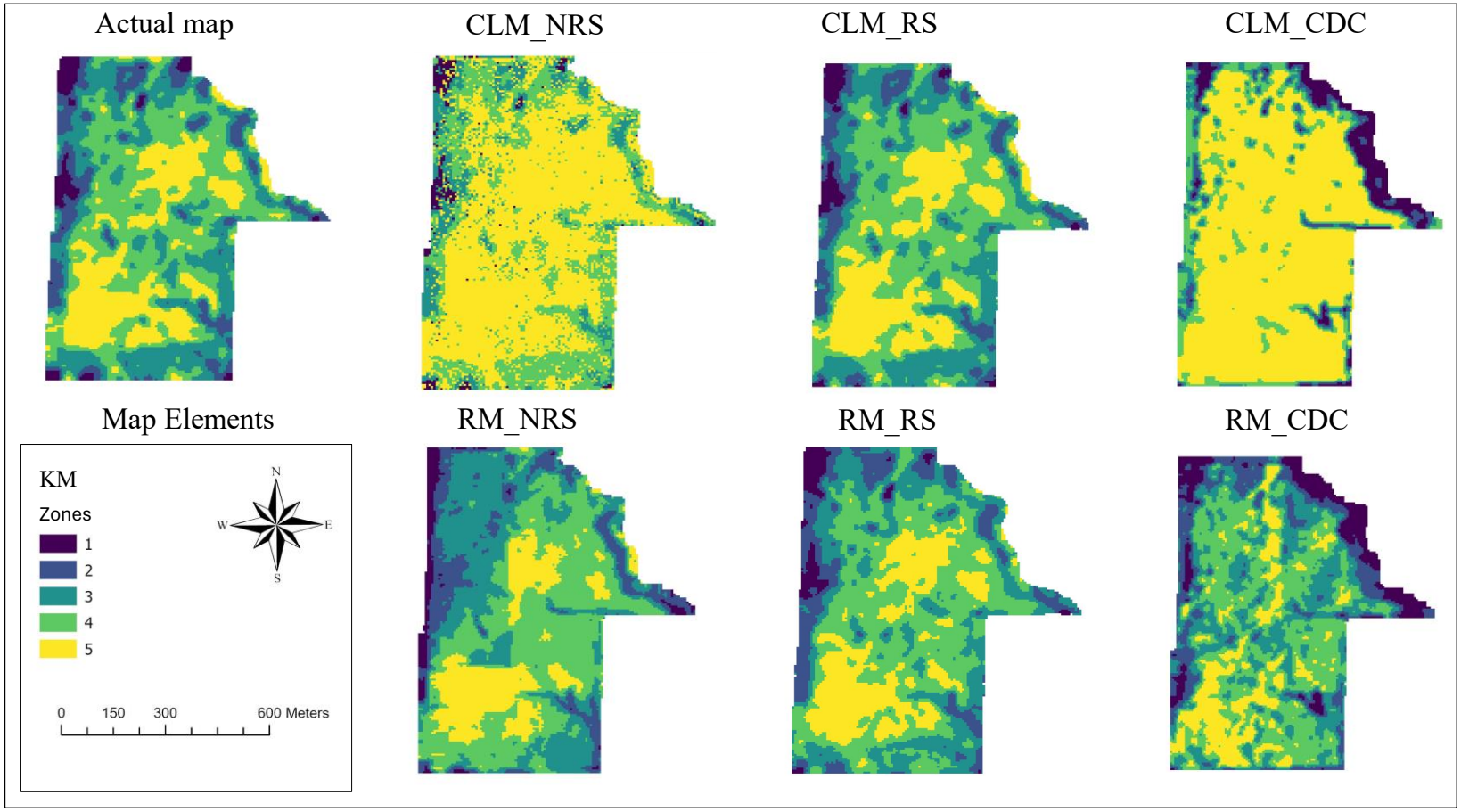

Figure 10. Actual and prediction maps for KM field.

CLM- Classification model RM – Regression model CDC- Combined data for Corn

NRS – Before adding remote sensing data RS – After adding remote sensing data

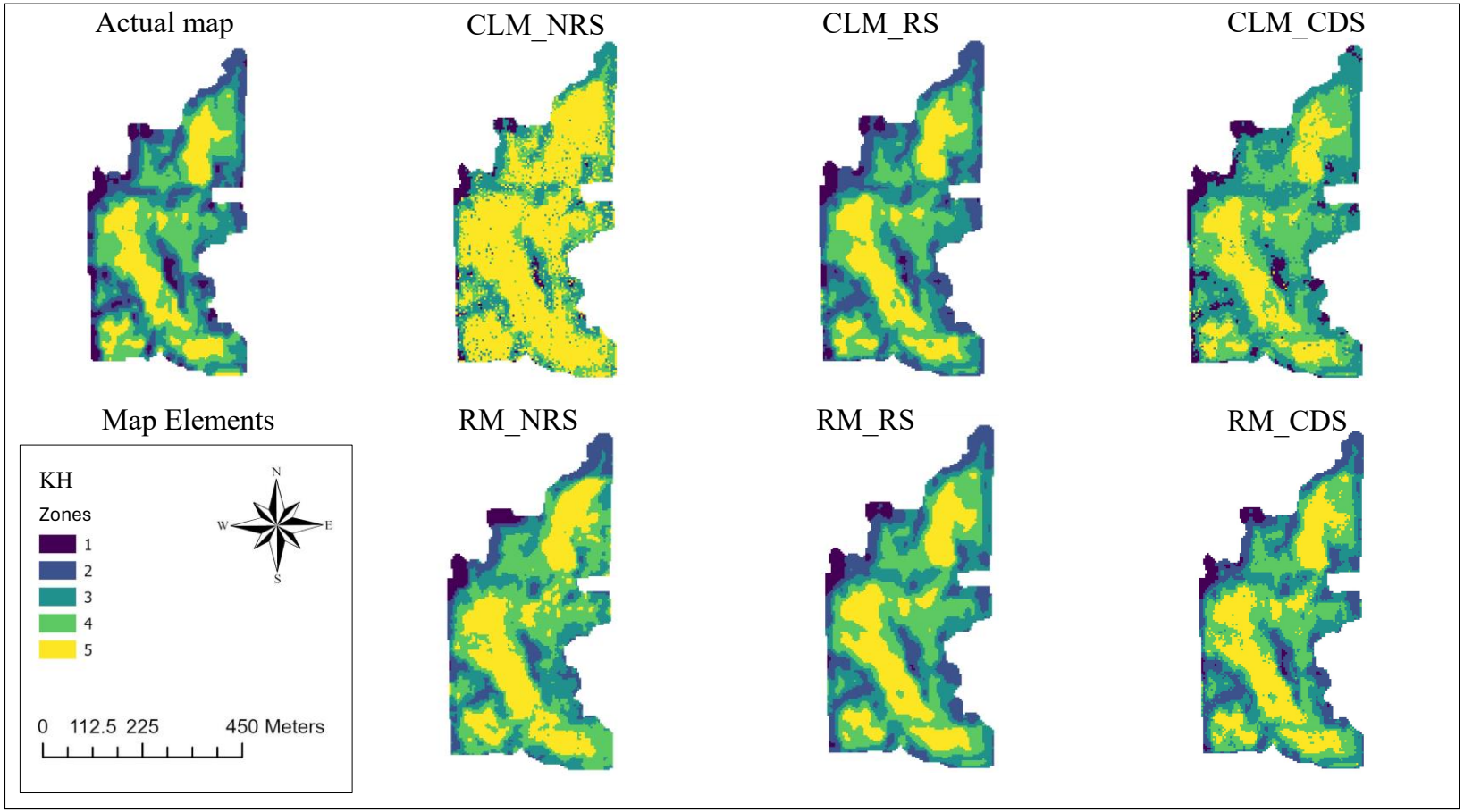

Figure 11. Actual and prediction maps for KH field.

CLM- Classification model RM – Regression model CDS- Combined data for Soybean

NRS – Before adding remote sensing data RS – After adding remote sensing data

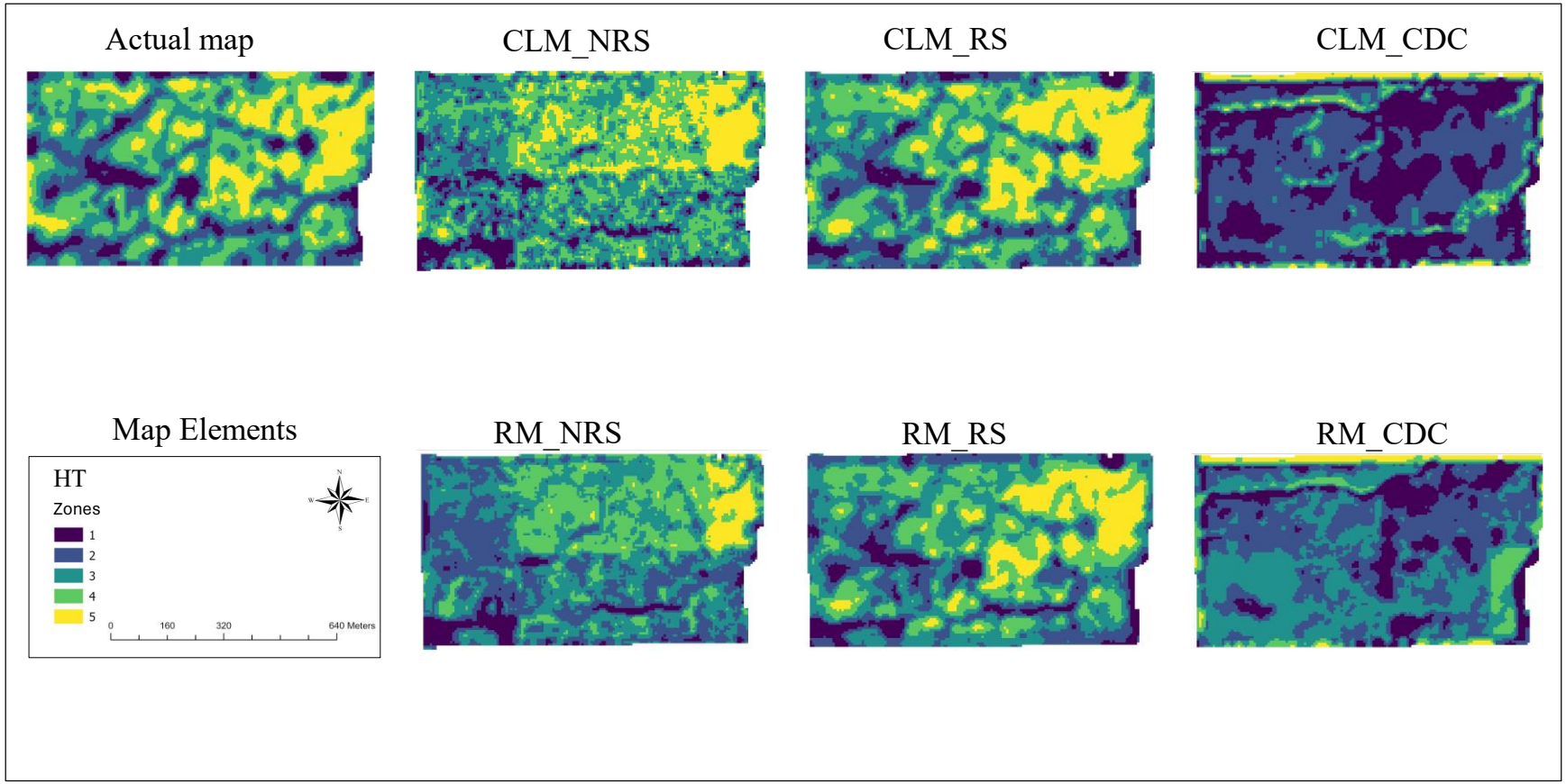

Figure 12. Actual and prediction maps for HT field.

CLM- Classification model RM – Regression model CDC- Combined data for Corn

NRS – Before adding remote sensing data RS – After adding remote sensing data

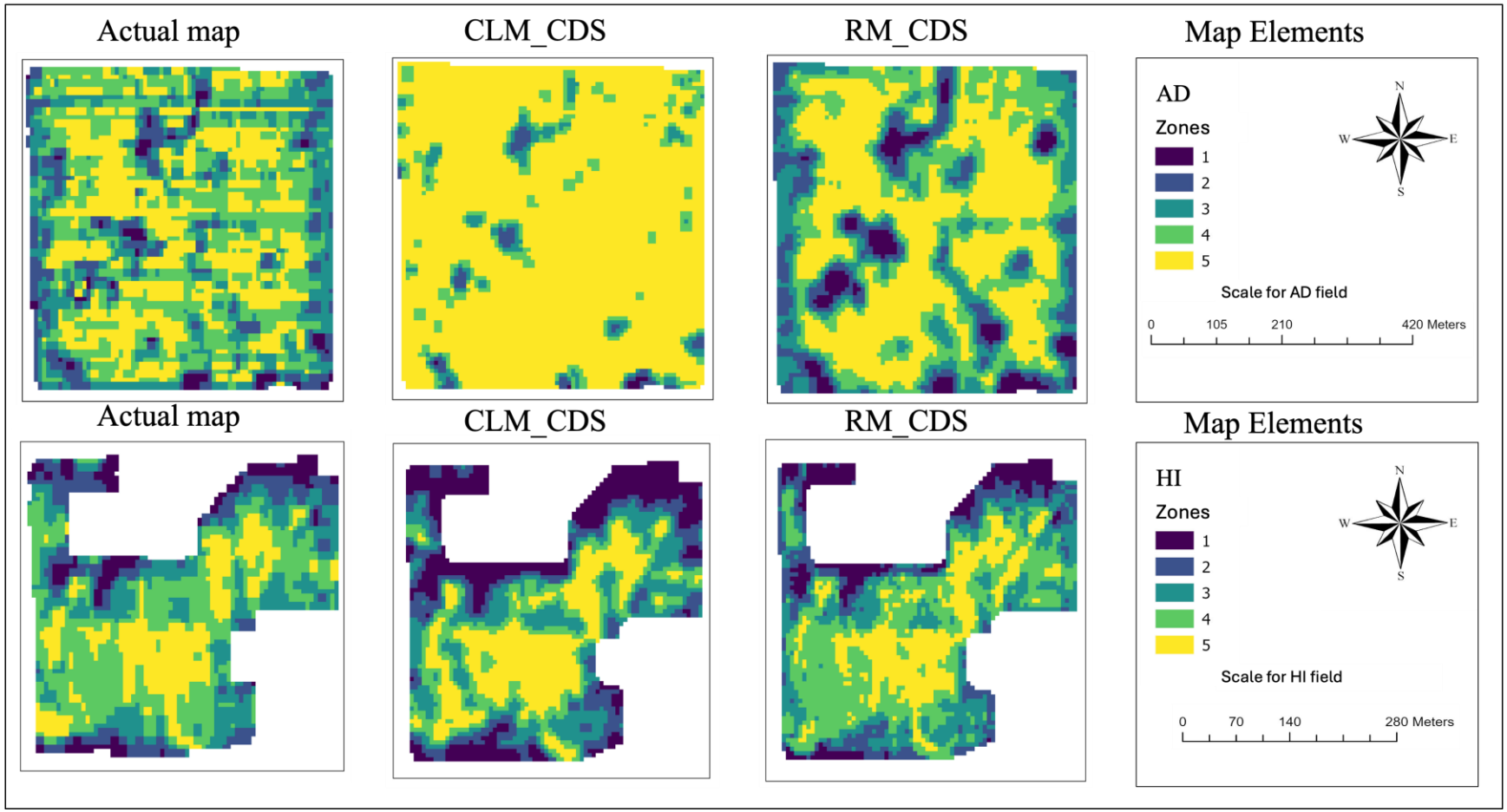

Figure 13. Actual and prediction maps for AD (top), HI (bottom) fields.
CLM- Classification model RM – Regression model CDS- Combined data for Soybean

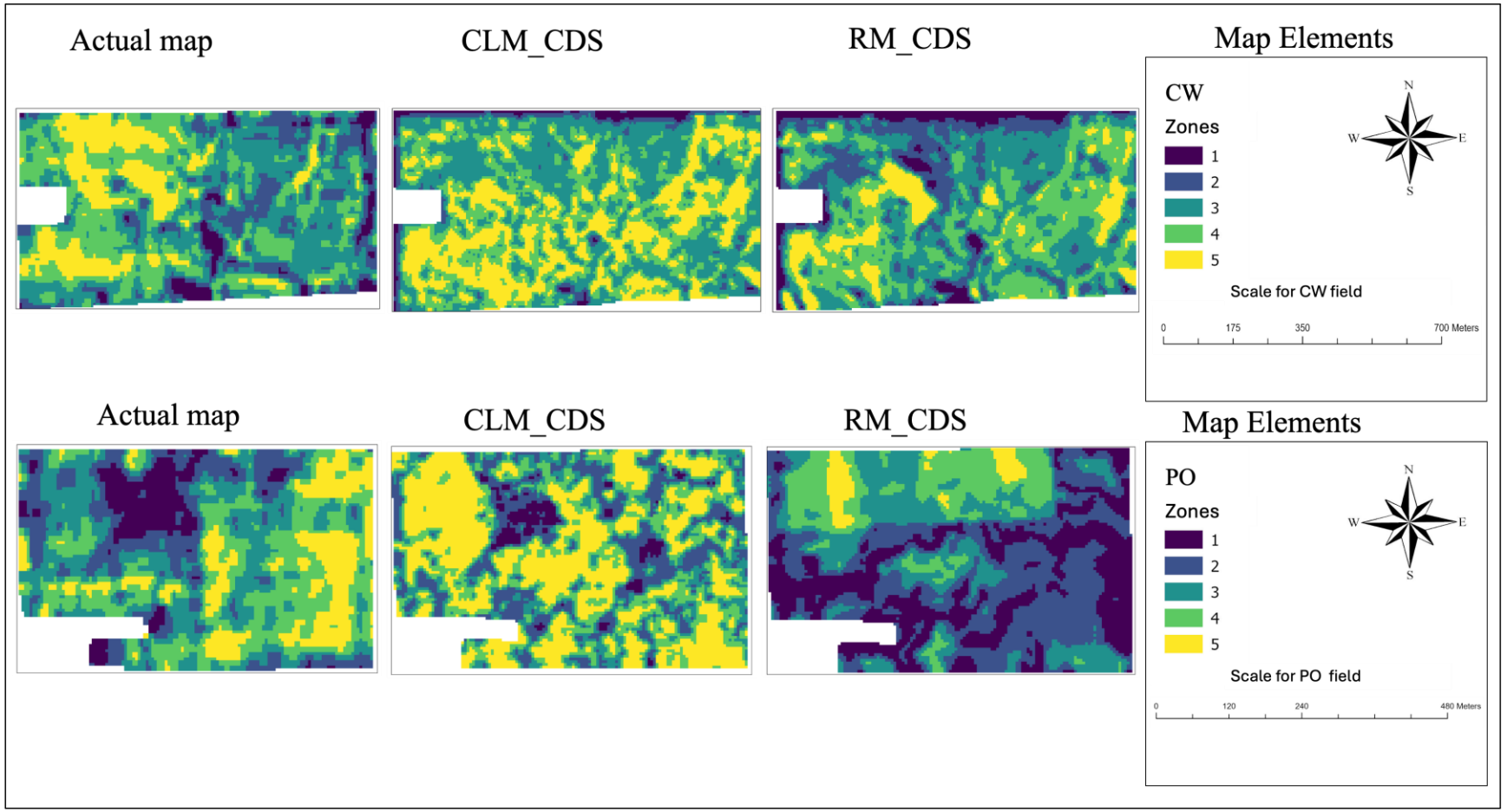

Figure 14. Actual and prediction maps for CW (top), PO (bottom) fields.
CLM- Classification model RM – Regression model CDS- Combined data for Soybean

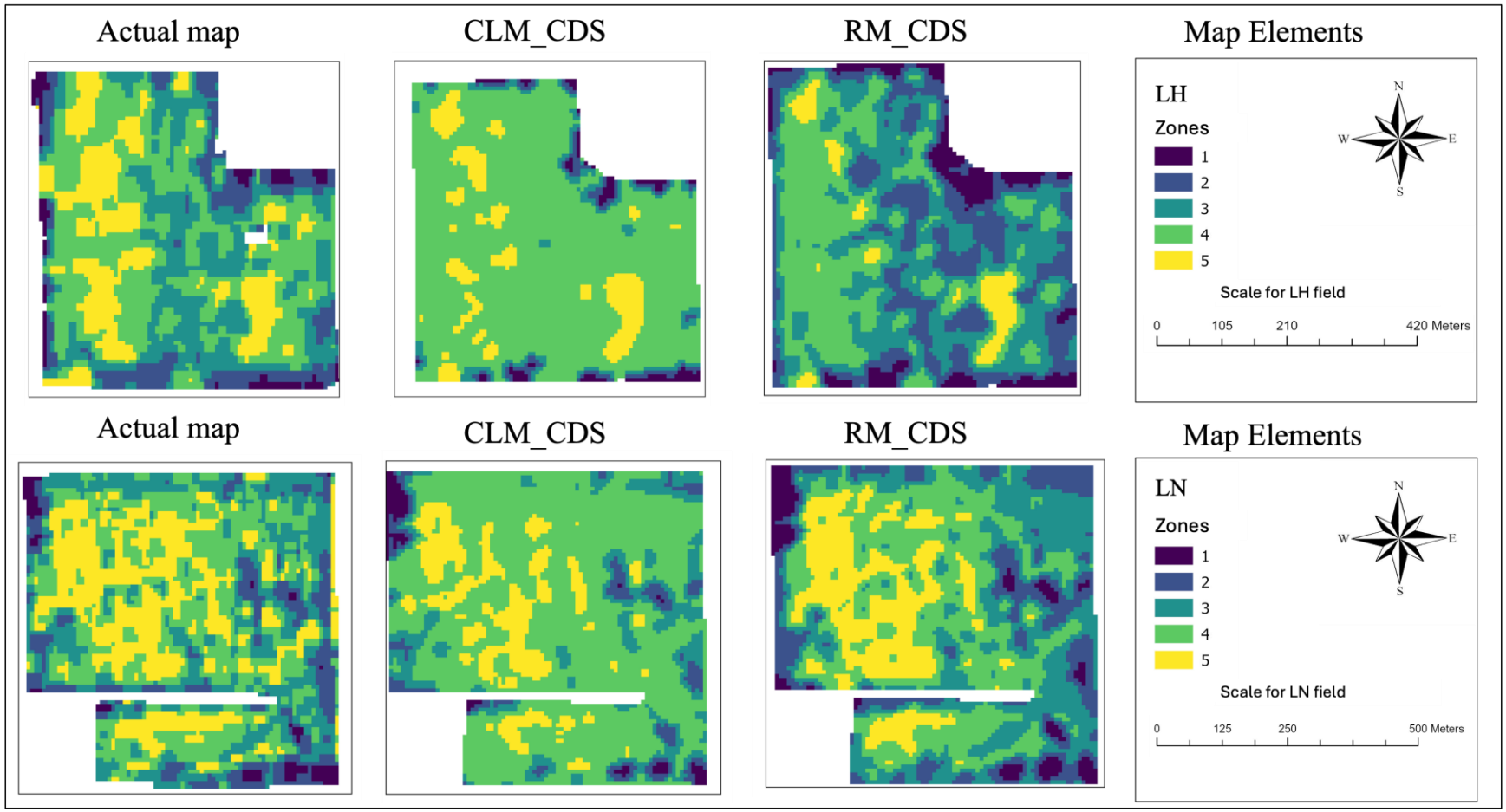

Figure 15. Actual and prediction maps for LH (top), LN (bottom) fields.
CLM- Classification model RM – Regression model CDS- Combined data for Soybean

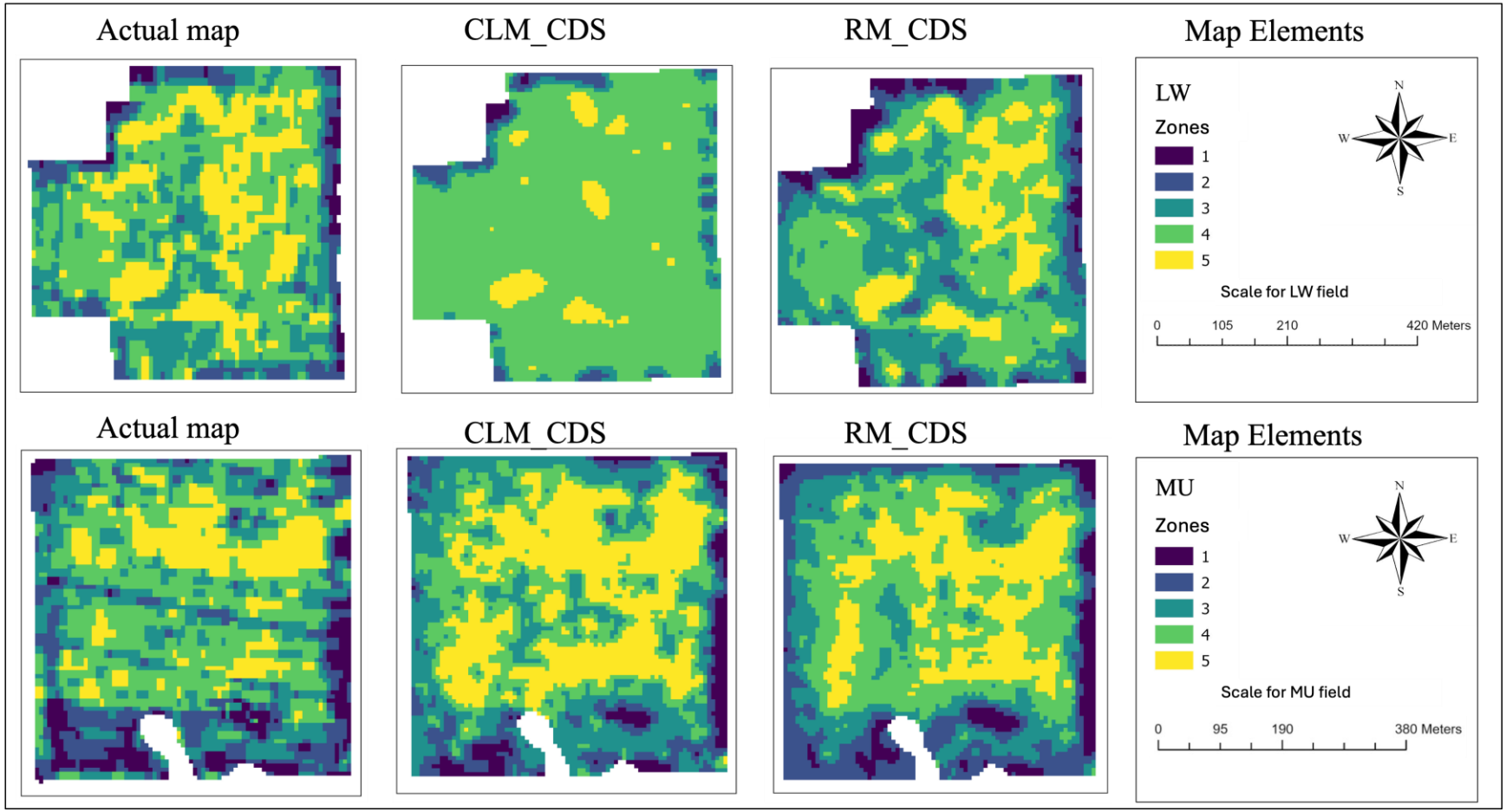

Figure 16. Actual and prediction maps for LW (top), MU (bottom) fields.
CLM- Classification model RM – Regression model CDS- Combined data for Soybean

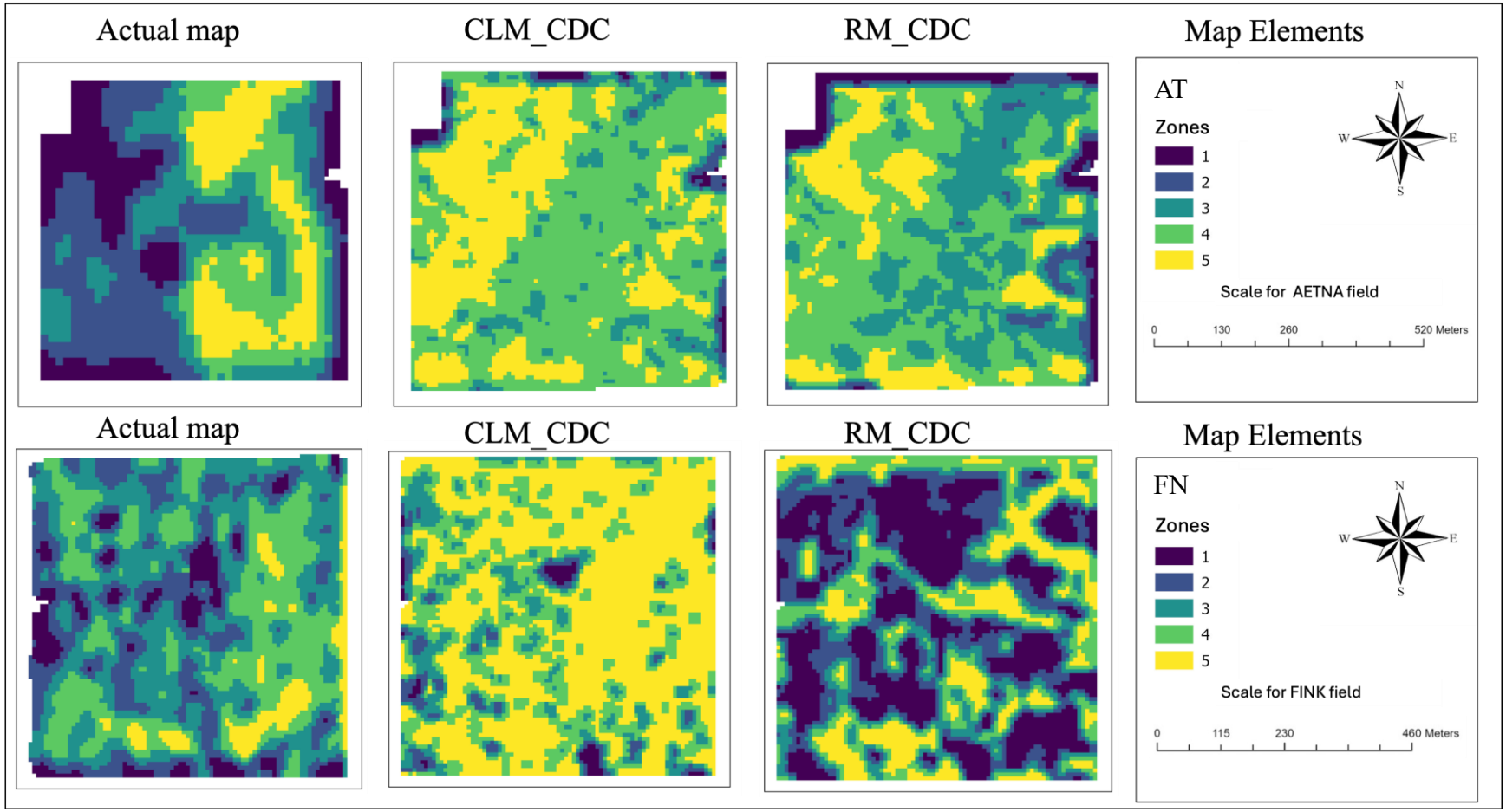

Figure 17.Actual and prediction maps for AT (top), FN (bottom) fields.
CLM- Classification model RM – Regression model CDC- Combined data for Corn

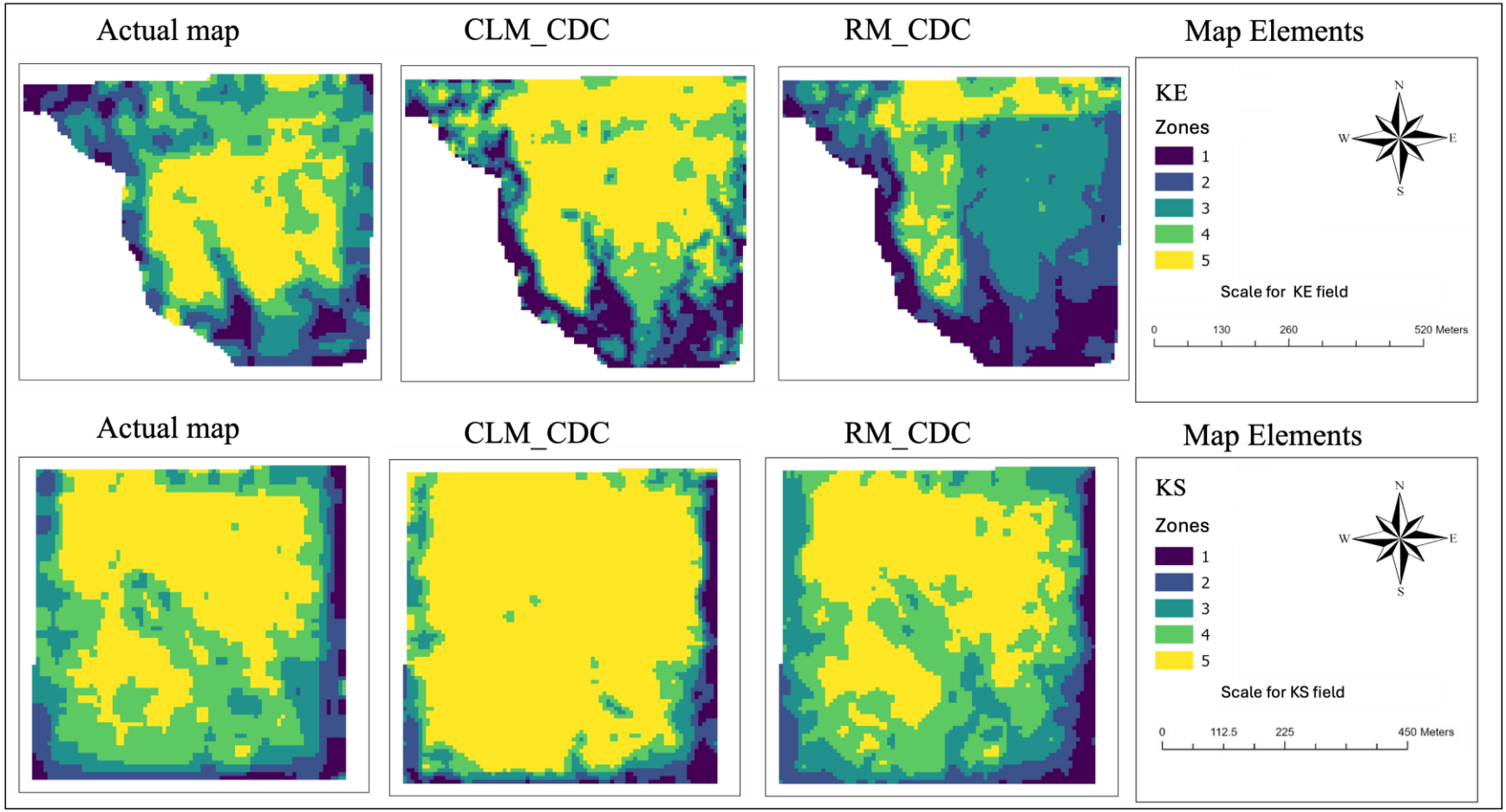

Figure 18. Actual and prediction maps for KE (top), KS (bottom) fields.
CLM- Classification model RM – Regression model CDC- Combined data for Corn

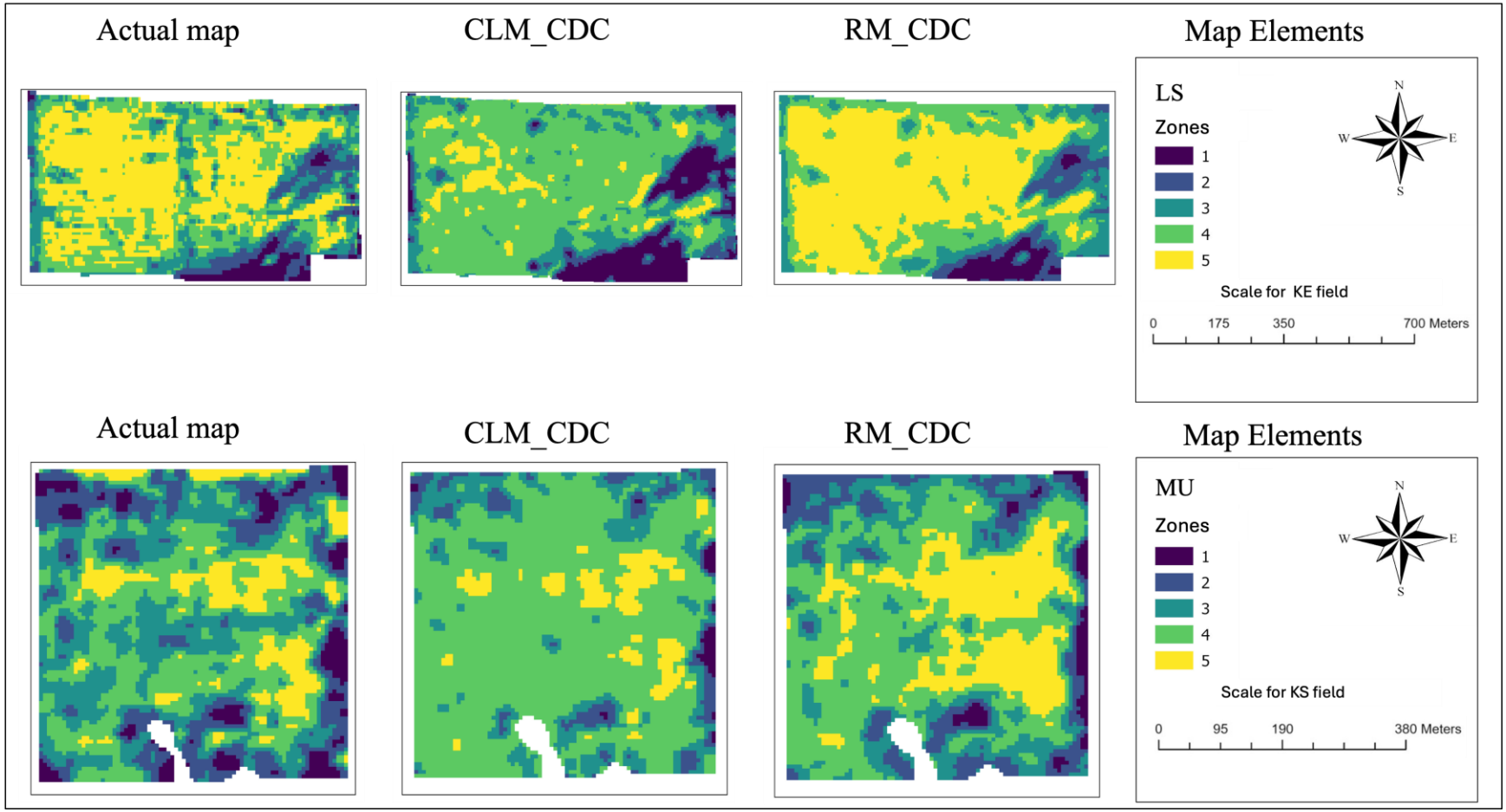

Figure 19. Actual and prediction maps for LS (top), MU (bottom) fields.
CLM- Classification model RM – Regression model CDC- Combined data for Corn