SOIL MOISTURE PREDICTION USING METEOROLOGICAL DATA, SATELLITE IMAGERY, AND MACHINE LEARNING IN THE RED RIVER VALLEY OF THE NORTH

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Umesh Acharya

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Soil Science

June 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

SOIL MOISTURE PREDICTION USING METEOROLOGICAL DATA, SATELLITE IMAGERY, AND MACHINE LEARNING IN THE RED RIVER VALLEY OF THE NORTH

By

Umesh Acharya

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Aaron L.M. Daigh
Chair

Dr. Frank Casey

Dr. Peter G. Oduor

Dr. Stephanie S. Day

Approved:

July 9, 2021
Date

Dr. Frank Casey
Department Chair

ABSTRACT

Weather stations provide key information related to soil moisture and have been used by farmers to decide various field operations. We first evaluated the discrepancies in soil moisture between a weather stations and nearby field; due to soil texture, crop residue cover, crop type, growth stage and duration of temporal dependency to recent rainfall and evaporation rates using regression analysis. The regression analysis showed strong relationship between soil moisture at the weather station and the nearby field at the late vegetative and early reproductive stages. The correlation thereafter declines at later growth stages for corn and wheat. We can adduce that the regression coefficient of soil moisture with four-day cumulative rainfall slightly increased with an increase in the crop residue resulting in a low root mean square error (RMSE) value. We then investigated the effectiveness of machine learning techniques such as random forest regression (RFR), boosted regression trees (BRT), support vector regression, and artificial neural network to predict soil moisture in nearby fields based on RMSE of a 30% validation dataset and to determine the relative importance of predictor variables. The RFR and BRT performed best over other machine learning algorithms based on the lower RMSE values of 0.045 and 0.048 $m^3 m^{-3}$, respectively. The Classification and Regression Trees (CART), RFR and BRT models showed soil moisture at nearby weather stations had the highest relative influence for moisture prediction, followed by the four-day cumulative rainfall and Potential Evapotranspiration (PET), and subsequently followed by bulk density and Saturated Hydraulic Conductivity (Ksat). We then evaluated the integration of weather station data, RFR machine learning, and remotely sensed satellite imagery to predict soil moisture in nearby fields. Soil moisture predicted with an RFR algorithm using OPtical TRApezoidal Model (OPTRAM) moisture values, rainfall, standardized precipitation index (SPI) and percent clay showed high goodness of fit (r^2=0.69)

and low RMSE (0.053 m^3 m^{-3}). This research shows that the integration of weather station data, machine learning, and remote sensing tools can be used to effectively predict soil moisture in the Red River Valley of the North among a large diversity of cropping systems.

ACKNOWLEDGMENTS

I express my sincere appreciation to my advisor and chair of supervisory committee Dr. Aaron L.M. Daigh for his continuous support, guidance, and the confidence he had in me to carry out these research projects. I also like to thank my advisory committee members, Dr. Frank Casey, Dr. Peter G. Oduor and Dr. Stephanie S. Day, for their advice and guidance in my research and studies. A special thanks to Dr. Jill Motschenbacher for her continuous support and motivation during my teaching and research.

I am grateful to Nate Derby for his assistance in image processing and soil sample collection from the field. I am also grateful to Megan and Zach for their field and laboratory assistance. A special thanks to Sudha GC Upadhaya from Washington State University for her help in writing machine learning algorithm and Pragati Singh (Remote Sensing Application Center Uttar Pradesh) for her help in writing codes in Google Earth Engine platform. I would also like to thank all the people in Department of Soil Science at the North Dakota State University for providing me with a friendly environment as well as knowledge throughout my studies.

Lastly, I am thankful to all my family members, Keshav Buba, Tara Mom, Usha, Uttam, Gokul, Bibek Uncle, Saru Aunti, Pragya (Nani), Prasanna (Cena), Ruby, Aaryan, Aaryadhya, Arjun, Sudha, and my friends Debankur, Pallavi, Tonoy, Tandrima, Rashad, Alec, Zach B., for believing in me and standing by my side, all the time, rendering their endless support and encouraging me in all possible ways.

DEDICATION

I dedicate my disquisition to my wife *Suprabha Khanal* and my parents *Bhesh Raj* and *Kalika Devi Acharya*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX TABLES

GENERAL INTRODUCTION

Soil moisture is an important variable in hydrology and climate studies due to its strong influence on water infiltration, runoff, evaporation, erosion and heat and gas fluxes (Verstraeten et al., 2007; Amani et al., 2017). Similarly, soil moisture plays a key role in farm activities such as crop selection and the timing of tilling, planting, applying fertilizers and harvesting (Hamman et al., 2002; Helms et al., 1996). However, the heterogeneity of soil moisture within and across spatial scales creates difficulties for both research efforts and land management decisions. The most accurate methods for representing soil moisture are point measurements (e.g., gravimetric sampling, in-situ electromagnetic sensors). Although, these methods are limited in terms of spatial extent, are time consuming and labor intensive (Brocca et al. 2007; Laguardia and Niemeyer 2008). Other methods with larger spatial extents include proximal and remote sensing technologies as well as hydrologic simulations to model soil moisture on the landscape (Babaeian et al., 2019). In contrast to point measurements, the larger spatial extents innately result in lower resolution and a loss of information in landscapes with complex physical attributes (e.g., topography, parent materials) and land management (e.g., crop rotations and diversity) and require adequate point-scale validation. Therefore, an efficient and reliable means to represent soil moisture in and across landscapes are highly desired by both the research and agricultural communities.

Researchers and farmers commonly use data from nearby weather stations to inform them on a location's soil moisture (if available), atmospheric conditions, and potential evapotranspiration. The key assumption for using these weather station data is that they adequately represent the actual conditions of nearby fields for some tasks of interest, even though these fields may differ in physical (e.g., soil texture, slope) and crop (e.g., type, previous

year's plant residues, growth stage) and attributes (Dalton et al., 2011; Rosenbaum et al., 2012). There are various factors such as crop type, soil texture, saturated hydraulic conductivity, topography, residue content affecting the soil moisture in crop field. The manual calculation of effect of these factors on the soil moisture is tedious work. There are two ways of predicting crop soil moisture using variables from weather station and crop factors; empirical model and machine learning methods. Empirical model uses statistical regression techniques to develop a mapping function based on the in-situ measurements of target variable and predictor variables. The advantage of empirical models is that they are typically fast to derive and do not require many inputs (Ali et al., 2015). The disadvantage of empirical models is the need for higher quality ground measurement that could be time consuming and expensive. Recently, the use of machine learning techniques has gained increased attention because it can overcome limitations of empirical and physical-based models. Popular machine learning methods currently in use are random forest, Artificial Neural Network (ANN), Support Vector Regression (SVR), Boosted Regression Trees (BRT), Classification and Regression Trees (CART), and Multiple Linear Regression (MLR). Matei et al. (2017) used different machine learning models for real time soil moisture prediction in Transylvania Depression of Romania. They used data (soil temperature, air temperature, precipitation) from a nearby weather station and used crop and soil information nearby station. Machine learning-based model (i.e., an RFR) achieved better performance when compared with the physics-based Richards equation model in predicting soil matric potential in the root zone (Gumiere et al., 2020).

Remote sensing has also been used as an advanced tool for agricultural interpretation since the early 2000s (Lillesand et al., 2008). A prime area of research in agriculture is the in-field variability of plant water stress across large scales, which directly relates to in-field soil

moisture (Bastiaanssen et al., 2000). Estimation of soil moisture provides farmers with key information of water stress, which aids in yield estimation, assessment of drought and excessive water conditions, and informs water management practices (e.g., irrigation, drainage) (Penuelas et al., 1993; Tucker, 1980). Remote sensing can be effectively used to estimate soil moisture because soil optical reflection and thermal emissions are highly correlated with soil moisture content (Zeng et al., 2016; Zhang and Zhou, 2016). In particular, the combination of remotely sensed visible and thermal infrared wavelengths provide more information for soil moisture estimation than either by themselves (Zhang et al., 2014). However, to get precise and accurate soil moisture estimations, both spatial and temporal information are needed (Zhang and Zhou, 2016). Remote sensing methods has provided tools for soil moisture mapping at large spatial and temporal scales (Das and Paul, 2015). Several mathematical models using remote sensing methods has been developed to estimate soil moisture using satellite optical image dataset such as Landsat or Sentinel that are freely available. Ali et al. (2015) and Paloscia et al. (2008) showed machine learning techniques (e.g., Artificial Neural Network, Support Vector Regression) can outperform other parametric approaches for estimating soil moisture and improved their performance with an increasing number of observed datasets. Therefore, integrating meteorological data from weather Mesonet and field characteristics (soil and crop data) along with OPTRAM soil moisture values in a machine learning algorithm has the potential to be a valuable tool in mapping high-resolution soil moisture across large areas.

The goal of this dissertation is to develop methodologies for producing accurate and high-resolution soil moisture maps throughout agriculturally dominated landscapes. In doing so, I focused on the Red River Valley of the North in North Dakota and Minnesota, USA, to develop, calibrate, and validate our methodologies. A variety of soil and crop characteristics,

meteorological data, and satellite imagery are used with machine learning methods to predict soil moisture and produce high-resolution maps.

This dissertation has three research chapters, where each will be prepared in the form of journal manuscripts. Each chapter title and specific objectives are:

- Chapter I: Factors affecting the use of weather stations data in predicting soil moisture for agricultural applications.
    - Determine the level of discrepancies in soil moisture between weather stations in the Red River Valley of the North (RRVN) and nearby agricultural fields.
    - Identify correlations of any discrepancies based on soil texture, crop type, residue cover and crop growth stage.
    - Determine the duration of temporal dependency of these soil moistures to recent rainfall and evapotranspiration rates.
- Chapter II: Machine learning for predicting field soil moisture using soil, crop, and nearby weather station data in Red River Valley of North.
    - Study the effectiveness of different machine learning tools in soil moisture prediction.
    - Find out the important predictor variables affecting field soil moisture content using machine learning tools.
- Chapter III: An integrated random forest–OPTRAM algorithm performed better than vegetation indices and OPTRAM for mapping surface soil moisture from Landsat 8 images.

- Obtain a representative surface soil moisture dataset across an agricultural geographic region with a complex mosaic of crop species and soil management on dates aligned with the Landsat 8 satellite.
- Calculate and determine the effectiveness of vegetation indices in predicting surface soil moisture.
- Predict surface soil moisture from the satellite images using OPTRAM.
- Evaluate if the OPTRAM predictions can be improved by incorporating weather station, soil, and crop data with a Random Forest machine learning algorithm.

References

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7(12), 16398–16421, doi:10.3390/rs71215841.

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., Dehnavi, S., 2017. Temperature-Vegetation-soil Moisture Dryness Index (TVMDI). Remote Sens. Environ. 197, 1-14. http://dx.doi.org/10.1016/j.rse.2017.05.026

Babaeian, E., Sadeghi, M., Jones, S.B., Montzka, C., Vereecken, H., Tuller, M., 2019. Ground, proximal, and satellite remote sensing of soil moisture. Rev. Geophys. 57(2), 530-616.

Bastiaanssen, W.G., Molden, D.J., Makin, I.W., 2000. Remote sensing for irrigated agriculture: Examples from research and possible applications. Agr. Water Manage. 46, 137-155.

Brocca L, Morbidelli, R., Melone, F., Moramarco, T., 2007. Soil moisture spatial variability in experimental areas of Central Italy. J. Hydrol. 333, 356–373. https://doi.org/10.1016/j.jhydrol.2006.09.004

Dalton, M., Andrews, P., Buss, P., Barrett, B., 2011. The use of the Evapotranspiration Stress Index (ETSI) to guide irrigation management in young olives. Acta Hortic. 924, 31-39 https://doi.org/10.17660/ActaHortic.2011.924.2

Das, K., Paul, P.K., 2015. Present status of soil moisture estimation by microwave remote sensing. Cogent Geoscience, 1(1), 1084669.

Gumiere, S.J., Camporese, M., Botto, A., Lafond, J.A., Paniconi, C., Gallichand, J., Rousseau, A.N., 2020. Machine Learning vs. Physics-Based Modeling for Real-Time Irrigation Management. Fro. Water, 2, 8. https://doi.org/10.3389/frwa.2020.00008

Hamman, B., Egil, D.B., Koning, G., 2002. Seed vigor, soilborne pathogens, pre-emergent growth, and soybean seeding emergence. Crop Sci. 42, 451-457. https://doi.org/10.2135/cropsci2002.0451

Helms, T.C., Deckard, E., Goos, R.J., Enz, J.W., 1996. Soybean seedling emergence influenced by days of soil water stress and soil temperature. Agron. J. 88, 657-661.

Laguardia, G., Niemeyer, S., 2008. On the comparison between the LISFLOOD modelled and the ERS/SCAT derived soil moisture estimates. Hydrol. Earth Syst. Sci., 12(6), 1339-1351.

Lillesand, T.M., Kiefer, R.W., Chipman, J.W., 2008. Remote sensing and image interpretation. Hoboken, NJ: John Wiley & Sons.

Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A data mining system for real time soil moisture prediction. Procedia Engineer. 181, 837-844.

Paloscia, S., Pampaloni, P., Pettinato, S., Santi, E., 2008. A comparison of algorithms for retrieving soil moisture from ENVISAT/ASAR images. IEEE T Geosci. Remote, 46(10), 3274-3284.

Penuelas, J., Filella, I., Biel, C., Serrano, L., Save, R., 1993. The reflectance at the 950– 970 mm region as an indicator of plant water status. Int. J. Remote Sens., 14, 1887-1905.

Rosenbaum, U., Bogena, H.R., Herbst, M., Huisman, J.A., Peterson, T.J., Weuthen, A., Western, A.W., Vereecken, H., 2012. Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale. Water Resour. Res., 48(10), 1-22.

Tucker, C.J., 1980. Remote sensing of leaf water content in the near infrared. Remote Sens. Environ., 10, 23-32.

Verstraeten, W.W., Verousraete, F., Feyen, J., 2007. Assessment of evapotranspiration and soil moisture content across different scales of observation. Sensors. 8, 70-117.

Zeng, W., Xu, C., Huang, J., Wu, J., Tuller, M., 2016. Predicting near-surface soil moisture content of saline soils from NIR reflectance spectra with a Modified Gaussian model. Soil Sci. Soc. Am. J., 80, 1496-1506. http://dx.doi.org/10.2136/sssaj2016.06.0188.

Zhang, D., Tang, R., Zhao, W., Tang, B., Wu, H., Shao, K., Li, Z.L., 2014. Surface soil water content estimation from thermal remote sensing based on the temporal variation of land surface temperature. Remote Sens., 6, 3170-3187.

Zhang, D., Zhou, G., 2016. Estimation of soil moisture from optical and thermal remote sensing: a review. Sensors 16(8), 1308

CHAPTER I. FACTORS AFFECTING THE USE OF WEATHER STATIONS DATA IN PREDICTING SOIL MOISTURE FOR AGRICULTURAL APPLICATIONS

Abstract

Weather stations often provide key information related to soil moisture, temperature, and evaporation. This information is used by farmers when deciding farm operations of nearby agricultural fields. However, the site conditions at the weather stations where data are recorded may not be similar with these nearby fields. The objective of this study was to determine the level of discrepancies in soil moisture between weather stations and nearby agricultural fields based on 1) the soil texture, crop residue cover, crop type, growth stages and 2) the duration of temporal dependency of soil moisture to recent rainfall and evaporation rates. Soil moisture from 25 weather stations in the North Dakota Agricultural Weather Network (NDAWN) and 75 nearby fields were measured biweekly during the 2019 growing season in Red River Valley of North. Field characteristics including soil texture, crop residue cover, crop type and growth stages along with rainfall and potential evapotranspiration were collected during the study period. The regression analysis between soil moisture at weather station and nearby crop field showed strong relationship at late vegetative and early reproductive stage then declined at later stage for corn and wheat, whereas correlation values increased for initial vegetative stage to podding for soybean. We can adduce that the regression coefficient of soil moisture with four-day cumulative rainfall slightly increased with an increase in the crop residue cover percentage resulting in a decreased Root Mean Square Error (RMSE). In general, we observed that soil moisture at weather stations could reasonably predict moisture in nearby agricultural fields considering crop type, soil type, weather, and distance from weather station.

Key words: soil moisture, weather station, Red River Valley, soil texture, rainfall, PET

Introduction

Soil moisture is an important variable in hydrology and climate studies due to its strong influence on water infiltration, runoff, evaporation, erosion and heat and gas fluxes (Verstraeten et al., 2007; Amani et al., 2017). Similarly, soil moisture plays key role in farm activities such as crop selection and the timing of tilling, planting, applying fertilizers and harvesting (Hamman et al., 2002; Helms et al., 1996). However, the heterogeneity of soil moisture within and across spatial scales creates difficulties for both research efforts and land management decisions. The most accurate methods for representing soil moisture are point measurements (e.g., gravimetric sampling, *in-situ* electromagnetic sensors). These methods are limited in terms of spatial extent and are time consuming and labor intensive (Brocca et al. 2007; Laguardia and Niemeyer 2008). Other methods with larger spatial extents include proximal and remote sensing technologies as well as hydrologic simulations to model soil moisture on the landscape (Babaeian et al., 2019). In contrast to point measurements, the larger spatial extents innately result in lower resolution and a loss of information in landscapes with complex physical attributes (e.g., topography, parent materials), land management (e.g., crop rotations and diversity) and require adequate point-scale validation. Therefore, an efficient and reliable means to represent soil moisture in and across landscapes are highly desired by both the research and agricultural communities.

Researchers and farmers commonly use data from nearby weather stations to inform them on a location's soil moisture (if available), atmospheric conditions, and potential evapotranspiration. The key assumption for using these weather station data is that they adequately represent the actual conditions of nearby fields for some tasks of interest, even though these fields may differ in physical (e.g., soil texture, slope) and crop (e.g., type, previous year's plant residues, growth stage) attributes (Dalton et al., 2011; Rosenbaum et al., 2012). In

the United States, there are 122 weather stations managed by National Weather Services to provide weather related products in addition to state-managed mesonets (NWS, 2020; NDAWN, 2020). The North Dakota Agricultural Weather Network (NDAWN) is an example of a state-managed mesonet, which provides up to 32 measured weather and soil parameters from 117 weather stations in North Dakota (83 stations), Minnesota (28 stations) and Montana (6 stations) (NDAWN, 2020). Similar state-level mesonets are also deployed in Kansas and Oklahoma (Kansas Mesonet, 2020; Oklahoma Mesonet, 2020).

In agricultural fields with annual grain crops, soil moisture is likely more dynamic over time than when under perennial cover. For instance, the amount, type and management (e.g., tillage) of crop residues left from previous growing season influences soil water evaporation and retention of soil moisture over time in the top soil (Dabney, 1998; Gwak and Kim, 2017). In addition, the live vegetation type and plant canopy cover modify the root-zone microclimate and affect evapotranspiration rates, while root morphologies and age strongly affect infiltration rates and patterns, and water uptake into the plant (Fernandez-Illescas et al., 2001). Therefore, the dynamics of live vegetation strongly affects soil moisture (Daigh et al., 2014; Thompson et al., 2010). Soil moisture measurement at weather stations is typically taken under a mowed perennial grass (i.e., turf), which starkly differs from the characteristics of nearby cropped fields (Patrignani and Ochsner, 2018). Moreover, if neighboring fields differ in soil texture, then soil moisture spatial variability and its dynamics over time will be impacted accordingly (Vereecken et al., 2007; Pan and Peters-Lidard, 2008; Ivanov et al., 2010; Vivoni et al., 2010). Using linear correlation and empirical orthogonal function analysis, Gwak and Kim (2017) reported that soil-particle-size distributions was a more dominating factor than vegetation in the soil moistures distribution. At larger scales, Dong and Ochsner (2018) reported that the variation of soil-

particle-size distributions across the landscape also controls soil moisture more than rainfall distributions during storm events.

The spatial extrapolation of measured soil and atmospheric conditions at weather stations is a major concern for representing nearby fields. However, most agricultural management decisions are also made based on inferences of what the conditions in those nearby fields will be in the following days or weeks. Weather forecasts of rainfall is likely the most obvious parameter to use for making such inferences. Rainfall history has a large impact on soil moisture and is a main determinant in farm activities (Western et al., 2002). Entekhabi and Rodriguez-Iturbe (1994) reported rainfall as the primary factor in controlling the state and subsequent evolution of soil moisture. Similarly, Pan et al. (2003) observed soil moisture to be a function of the time-weighted average of previous cumulative rainfall over a period of 14 days. However, evapotranspiration, air and soil temperatures, and wind speeds are also some of the more widely used weather data from stations to make inferences on near-future soil moisture conditions (Western et al., 2002). The variety of factors influencing soil moisture variability in space and time (e.g., soil physical properties, topography, microclimate, groundwater, evapotranspiration) presents a barrier for farmers and agricultural consultants to infer the representativeness of weather station data to nearby fields readily and efficiently (Famiglietti et al., 1998; McMillan and Srinivasan, 2015; Rosenbaum et al., 2012; Vereecken et al., 2007; Western et al., 2002).

Therefore, it is important to determine discrepancies in soil moisture between local weather stations and nearby agricultural fields. Moreover, identifying correlations of any discrepancies to differences in soil type, residue cover, or crop type and growth stage can then guide the development of simple quantitative relationships to extend weather station data to inform on-farm management decisions. Such discrepancies are intuitively expected. However,

little to no evidence is currently reported in the literature. To our knowledge, the literature lacks any such evaluations for the upper interior plains of North America. Thus, our objectives are to: (i) determine the level of discrepancies in soil moisture between weather stations in the Red River Valley of the North (RRVN) and nearby agricultural fields, (ii) identify correlations of any discrepancies based on soil texture, crop type, residue cover and crop growth stage, and (iii) determine the duration of temporal dependency of these soil moistures to recent rainfall and evapotranspiration rates.

Material and Methods

Study area, weather station network and data collection

The study area was in North Dakota and Minnesota within the Red River of North (RRVN). The Red River of North extends 885 km northward from its source near Breckenridge, Minnesota in the United States (US) to Lake Winnipeg in Canada. The segment of river in the US (634 km) forms most of the border between Minnesota and North Dakota. The Red River valley of north is a glaciolacustrine lake bed formed by the ancient Lake Agassiz, which existed for about 4,000 years. The topography is minimal with a gradient of only 1:5000 (1 meter per 5 kilometer). The dominant soil orders in RRVN are Mollisols and Vertisols, whereas soil texture ranges from loamy sand to clay. The large range in textures can be attributed to variations in the lake deposits and formation of braided streams as the ancient lake drained to the north in around 8,000 years ago. The parent material is poorly drained and consists of gray, slickensided, flat clays of Brenna/Argusville formations, which are overlain by the tan-buff, laminated silty clays of the Sherack Formation. Shales within the parent materials commonly result in the shallow perched water tables being saline or saline-sodic. The major crops grown in this region are (*Zea mays* L.), soybean (*Glycine max* (L.) Merr.), wheat (*Triticum aestivum* L.), barley (*Hordeum*

vulgare L.), sugar beet (*Beta vulgaris*) along with canola (*Brassica napus*), sunflower (*Helianthus annuus* L.), potato (*Solanum tuberosum* L.), dry beans (*Phaseolus vulgaris*), and oats (*Avena sativa* L.) as the major crops grown annually. Summers are long and warm, whereas winters are frigid, snowy, windy and partly cloudy year-round. The average annual air temperature is 4 °C, typically varies from -16 °C to 29 °C and rarely below -27 °C or above 32 °C, whereas, 30-year mean annual rainfall is 60 cm and snowfall of 317 cm (NOAA/NCEI, 2020).

The North Dakota Agricultural Weather Network (NDAWN) was used for the study. NDAWN reports 32 weather parameters (e.g. air temperature, rainfall, wind direction, soil moisture) at 117 weather stations, which includes stations in North Dakota (N = 83), Minnesota (N = 28) and Montana (N = 6). A subset of these stations (i.e., those located in the RRVN) were selected for this study. This included a total of 25 stations, where 15 stations were located across 8 counties in North Dakota and 10 stations were located across 7 counties in Minnesota (Figure 1.1).

Weather station data and measurements in nearby agricultural fields of the study area were collected during the cropping season from June 1 to September 30 in 2019. Three nearby agriculture fields (corn, soybean, wheat, sugarbeet, potato, dry bean, canola) within the range of 30 meter to 2 kilometers were selected around weather station (N = 75 fields). From each field, three different soil samples were randomly selected to determine soil moisture content. Soil samples were collected in 16-day intervals from the field and weather station between June to September 2019 (Table 1.1).

Figure 1.1. Map showing counties of North Dakota and Minnesota and weather stations under study area around Red River Valley. Black dots in map represents weather stations and italic with underline word represents counties

Table 1.1. Soil sampling date with corresponding weather station for soil moisture determination for year 2019

Weather station	Date Sampled (2019)
Campbell, Mooreton, Wahpeton	6/27, 7/13, 7/29, 8/14
Leonard, Sabin, Fargo, Ulen, Prosper, Galesburg, Perely, Hillsboro, Ada, Waukon, Mayville, Finley, Eldred, Grand Forks, Forest River, Inkster, Warren, Grafton, St. Thomas, Kennedy, Cavalier, Humboldt	6/18, 7/20, 8/5, 8/21
Grafton, St. Thomas, Kennedy, Cavalier, Humboldt	7/27

Determination of soil moisture

Soil moisture was measured using the gravimetric method for each location and sample date (N = 985). NDAWN has only recently started installing soil moisture sensors in the top soil

(0-6 cm) at their weather stations, and only 4 stations under our study had these sensors at the time of the study. Therefore, soil samples were collected from all weather stations to determine soil moisture. GPS coordinates for each station and sampling location were also recorded. Three composite soil samples (0-6 cm) from each field and station were collected using Uhland cores to determine soil moisture. Soil was sampled using core sampler with dimension (6 cm x 8 cm), the field-wet weight of the soil was recorded, and then oven dried at 105°C for 48 hours. The weight of dry soil was again recorded, and gravimetric water content was determined as the mass of water lost due to drying. The soil's volumetric water content (VWC) was calculated by multiplying gravimetric water content with the soil bulk density (e.g. Reynolds, 1970).

Crop type and growth stages

The major crops grown in RRVN are corn, soybean, wheat, barley, sugar beet along with canola, sunflower, potato, dry beans, and oats. For this study, the selected fields nearby the weather stations, were planted with soybean (N=24), wheat (N=18), corn (N=16), sugar beet (N=6), drybeans (N=5), oats (N=2), barley (N=1), potato (N=1), canola (N=1), and alfalfa (N=1). Soil samples taken after crops were planted and germinated. Growth stages for each crop were recorded every 16 days throughout the growing period until harvest. The growth stages for each crop were determined using standards developed by the United States Department of Agriculture (e.g. USDA, 2020). These coincided with the dates for soil sampling, and soil moisture values determined for each growth stage.

Antecedent site characteristics: Crop residue cover, soil texture and saturated hydraulic conductivity (Ksat)

Crop residue cover was determined along eight transects per sample site using the rope method (i.e., residue presence at 100 points along 15 m oriented 45° to plant rows) (Daigh et al.

2019) at the start of growing season. Crop residue was then pooled into three categories: <10, 20-30 and 50-60% crop residue cover. Soil texture was determined for each site (i.e., weather stations and nearby fields) using the pipette method described by Gee & Bauder (1986) using composite soil samples. Saturated hydraulic conductivity (Ksat) was estimated using the Rosetta neural network pedotransfer function in Hydrus-1D, which used input data of sand (%), silt (%), clay (%), bulk density (g cm^{-3}), and water contents at 33 and 1500 kPa suctions (cm^3 cm^{-3}) (www.pc-progress.com) (Schaap et al., 2001; Simunek et al., 2008). Water contents at 33 and 1500 kPa suctions were determined using pressure plate apparatus (e.g. Richards, 1948).

Rainfall and potential evapotranspiration

Rainfall at the NDAWN stations was measured hourly at a 1m height above the soil surface using TE525 tipping bucket rain gauges (Texas Electronics TR-525I, Dallas, Texas). Each bucket tip measures 0.254 millimeters of rainfall. The Potential Evapotranspiration (PET) estimates of the maximum daily crop water loss when water is readily available. PET is calculated from solar radiation, dew point temperature, wind speed, and air temperature using the Penman (Penman, 1948) equation and is based on alfalfa which is called reference ET. Rainfall and PET (mm) for the preceding 10 days before soil sampling were downloaded from each weather station (https://ndawn.ndsu.nodak.edu/) and used to calculate cumulative values.

Statistical analysis

Linear and nonlinear regression was performed and Pearson correlation coefficients determined to describe the relationships between soil moisture at the weather stations (independent variable) and the nearby cropped fields (dependent variable) using Proc Reg in SAS software (version 9.4, SAS, 2017). The analysis was repeated by pooling the data for each factor [i.e., crop type, crop growth stage, crop residue cover, soil texture, distance from weather

station, and their interactions] separately. The recent cumulative rainfall and potential evapotranspiration history at the weather station were compared with the current soil moisture using non-linear regression using Proc Reg in SAS software (version 9.4, SAS, 2017). The analysis was repeated by pooling the data for each factor [i.e., crop type, crop growth stage, crop residue cover, soil texture, distance from weather station, and their interactions] separately. The regression parameters (slopes and intercept), correlation coefficients, and root mean square error (RMSE) are reported and discussed below.

Results

VWC discrepancies between crop field and weather station

Soil moisture ranged from 0.028 to 0.523 $m^3\ m^{-3}$ across the study area of the RRVN and sampling time frame (May to September, 2019). VWC at weather stations and nearby agricultural fields were linearly correlated, with the weather station VWC explaining 50% (r^2=0.50, N=675, slope=0.7, RMSE=0.0654 $m^3\ m^{-3}$) of the variance for the nearby fields (Figure 1.2). Distances to 2 Km from the weather station moderately affected this relationship (Appendix A1). The correlation coefficient was higher (r^2=0.55, N=215) for fields nearer (0-100m) as compared to fields farther (1200-2000m) from weather stations (r^2=0.40, N=42).

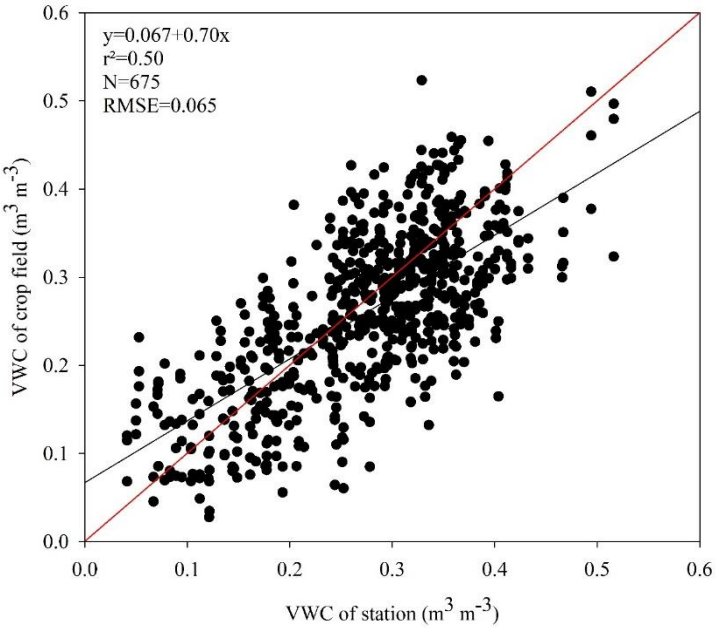

Figure 1.2. Linear relationship between volumetric water content (VWC) of crop fields with nearby weather stations in the Red River Valley during 2019.

VWC discrepancy due to crop type and their growth stage

Discrepancies associated with crop types and growth stage were apparent between VWC at the weather stations and nearby fields (Figure 1.3, Appendix A2). Fields planted to dry beans (r^2=0.69, N=33, RMSE=0.041 m^3 m^{-3}) had the highest correlation, followed by wheat (r^2=0.56, N=159, RMSE=0.06 m^3 m^{-3}) and corn (r^2=0.46, N=156, RMSE=0.068 m^3 m^{-3}), whereas, the lowest correlations were in sunflower (r^2=0.41, N=9, RMSE=0.061 m^3 m^{-3}) and barley (r^2=0.18, N=9, RMSE=0.052 m^3 m^{-3}), which also had the lowest sample size. Moreover, the slope of all the linear regression equations was less than 1 (i.e., corn = 0.81; wheat = 0.72; sugarbeet = 0.70; soybean = 0.69; alfalfa = 0.45) (Appendix A1).

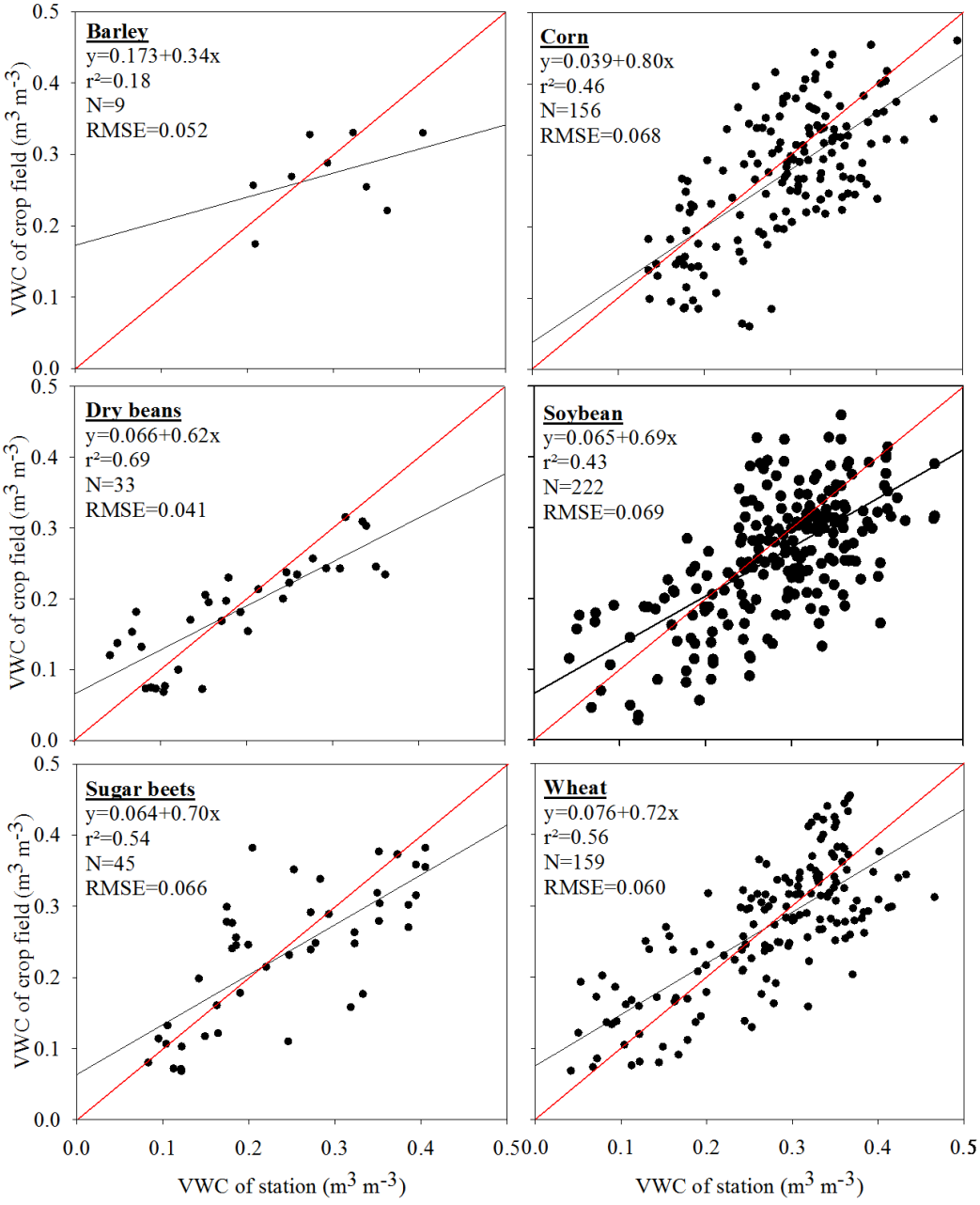

Figure 1.3. Linear relationship between Volumetric Water Content (VWC) of crop fields with nearby weather stations under different crop types.

Regression coefficients increased with corn growth stage [V10 stage (r^2=0.92), V11 stage (r^2=0.99)] until the silking reproductive phase and then declined [tasseling (r^2=0.59), grain filling (r^2=0.78)]. Wheat expressed a similar trend [tillering (r^2=0.17); flowering (r^2=0.68); hard dough

(r^2=0.590 and after harvest (r^2=0.24)], whereas correlations in soybean continued to increase [V1 stage (r^2=0.11); V2 stage (r^2=0.15); flowering (r^2=0.51); podding (r^2=0.70)].

VWC discrepancy due to residue cover and soil texture

Crop residue cover and soil texture had a moderate influence on the disparity between the weather stations and nearby fields. Crop fields with the lowest amount of crop residue cover (<10%) had the highest correlation (r^2=0.63, N=275, RMSE=0.058 m^3 m^{-3}) with the VWC of weather station as compared to higher residue covered fields [20-30% residue (r^2=0.44, N=198, RMSE=0.066 m^3 m^{-3}); 50-60% residue (r^2=0.46, N=198, RMSE=0.067 m^3 m^{-3})] (Figure 1.4). Soils with a relatively high clay content, such as clay (r^2=0.63, N=48, RMSE=0.061 m^3 m^{-3}), clay loam (r^2=0.57, N=69, RMSE=0.063 m^3 m^{-3}), silty clay loam (r^2=0.52, N=117, RMSE=0.054 m^3 m^{-3}) and silty clay (r^2=0.46, N=153, RMSE=0.073), had higher correlation coefficients as compare to soils with a high sand content (Figure 1.5).

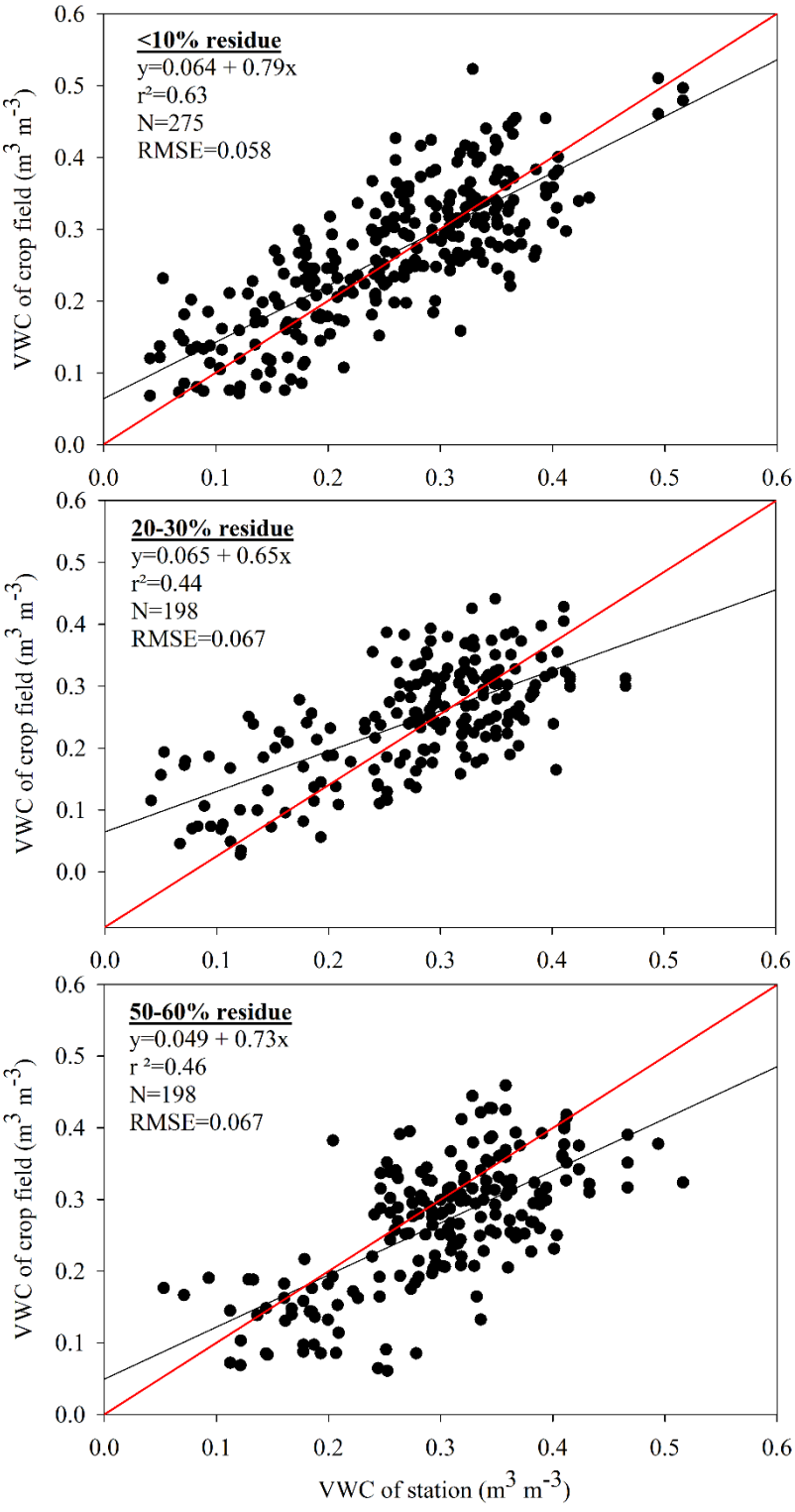

Figure 1.4. Linear relationship between volumetric water content (VWC) of crop field with weather stations at different residue percentage (<10, 20-30, 50-60).

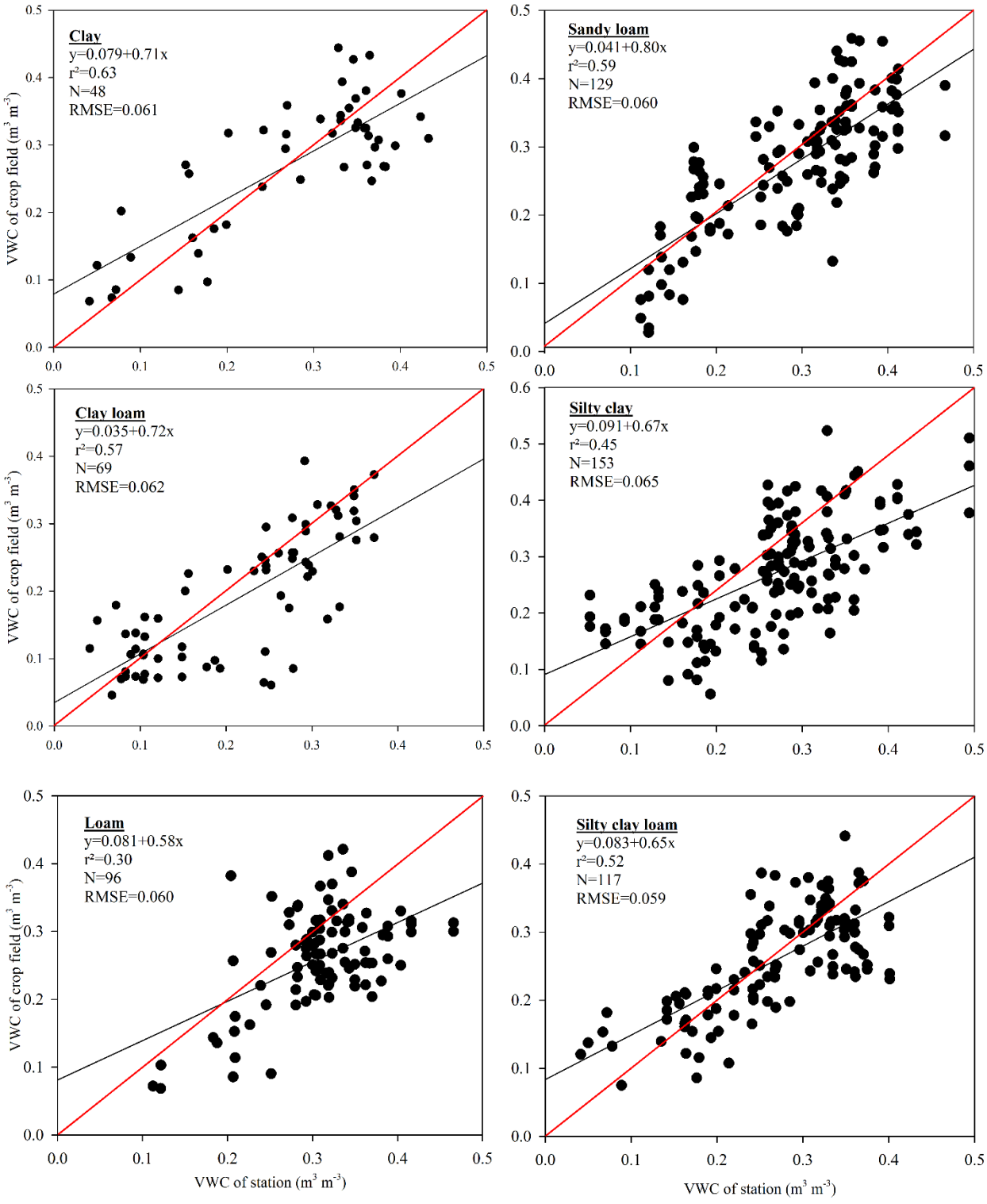

Figure 1.5. Linear relationship between Volumetric Water Content (VWC) of crop field with weather station at different type of soil texture in the study area.

Temporal dependency of soil moisture in crop fields to recent rainfall and PET rates

Soil moisture expressed a non-linear (cubical) relationship with past cumulative rainfall and PET (5 days) measured from the weather station. The highest correlation (r^2=0.49) was observed between soil moisture and a four-day cumulative rainfall that was improved significantly from a one-day cumulative rainfall (r^2=0.16). Similarly, the highest correlation (r^2=0.29) was observed between soil moisture and a four-day cumulative PET (Appendix A3).

The non-linear relationship between soil moisture with four-day cumulative rainfall had various weak to strong influences by crop residue cover, crop type, distance from weather station and soil texture (Figure 1.6; Appendix A4). The regression coefficient of soil moisture with four-day cumulative rainfall slightly increased with an increase in the crop residue cover percentage (<10, 20-30, 50-60%) from 0.48 to 0.51 and RMSE decreased from 0.068 to 0.063 m^3 m^{-3}. The highest correlation coefficient between soil moisture and a four-day cumulative rainfall was observed with alfalfa (r^2=0.93), followed by oats (r^2=0.86), sugarbeet (r^2=0.71), dry beans (r^2=0.65), wheat (r^2=0.56), corn (r^2=0.48) and lowest in soybean (r^2=0.45). The non-linear relationship between soil moisture and four-day cumulative rainfall shows crop fields near to weather station (100-200m) had higher correlation coefficient (r^2=0.65), whereas fields further away (1200-2000m) had a lower coefficient (r^2=0.25). A strong non-linear relationship was observed between soil moisture and four-day cumulative rainfall for soils having high clay content [clay (r^2=0.75), silty clay loam (r^2=0.65), clay loam (r^2=0.52)], whereas a weak relationship was observed with soils having high sand content.

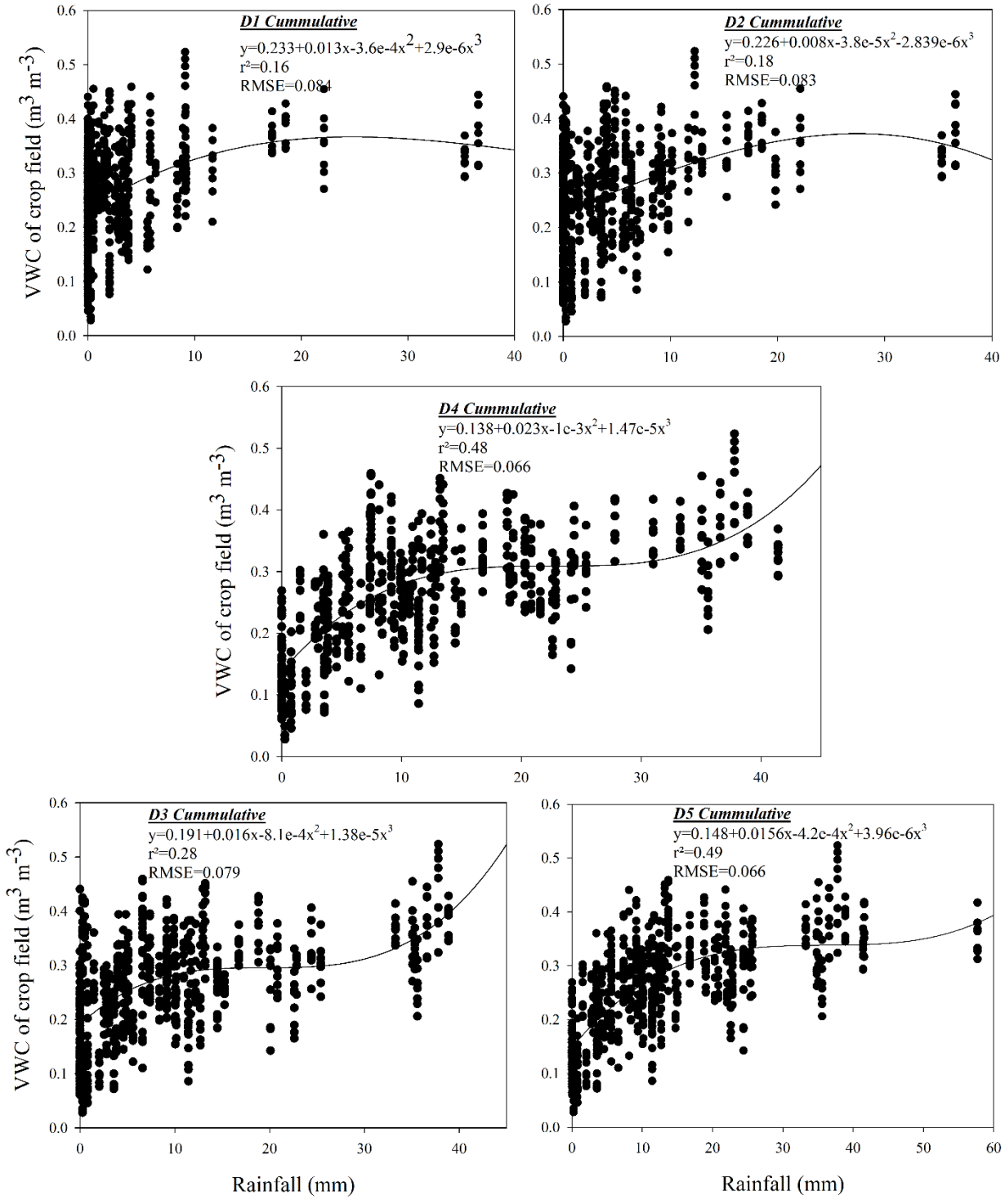

Figure 1.6. Non-linear relationship between volumetric water content (VWC) of crop field (N=675) with cumulative rainfall for past five days (D1, D2, D3, D4 and D5) for the study area.

Similarly, the highest correlation coefficient of soil moisture with four-day cumulative PET (r^2=0.37) was observed with 20-30% crop residue cover followed by the 50-60% (r^2=0.31) and lowest observed with <10%. The RMSE was also lowest with 20-30% crop residue cover (0.071 m^3 m^{-3}) followed by 50-60% (0.076 m^3 m^{-3}) and <10% (0.083 m^3 m^{-3}) (Appendix A5). For different types of crops, similar trends were evident between soil moisture and a four-day PET as with four-day cumulative rainfall. In contrast to cumulative rainfall, the opposite was observed between soil moisture and four-day cumulative PET, where farther crop fields (800-1200m) had higher correlation coefficients (r^2=0.57) and the nearest fields (0-100m) had a lower coefficient (r^2=0.33).

Similarly, correlation between soil moisture and four-day cumulative PET for soils was higher with high clay content [clay (r^2=0.59), clay loam (r^2=0.59), silty clay (r^2=0.51)] as compared to soils having high sand content [loamy sand (r^2=0.09)] as with cumulative rainfall.

Discussion

In general, we observed that soil moisture at weather stations could reasonably predict moisture in nearby agricultural fields (Figure 1.2) considering crop, soil, weather, and distance from weather station. This corroborates with findings by Famiglietti et al. (1998) regarding correlations between topographical attributes, soil properties and soil moisture measured along distances of 200 m. Therefore, the discrepancies in soil moisture observed in the present study are likely due to spatial heterogeneities of soil characteristics (Hu et al., 1997), vegetation characteristics (Qiu et al., 2001), and land management practices (Daigh et al., 2018).

As expected, the analysis showed the moisture prediction weakens with an increase in distance from the weather stations. This is likely due to change in soil moisture spatial patterns caused by the field variations in the landscape as well as other autocorrelated factors (soil

texture, vegetation, rainfall, evapotranspiration) that influence local hydrologic processes (Bardossy and Lehmann, 1998; Famiglietti et al., 1998; Western et al., 1999, Brocca et al., 2007). The large spatial and temporal variability of the study area might have resulted in the lower prediction value which can be improved by considering those factors in a prediction model (McMillan and Srinivasan, 2015). The changes in the spatial pattern of soil moisture was studied by Hawley et al. (1983) in the flat areas of Central Italy under different soil wetness condition, Cunningham et al. (1978) in ten-year long revegetating study in Australia, and Dunin and Reyenga (1978) in evaporation study of sub-humid grassland of Australia.

Crop type and their growth stages showed weak to strong relationships in the soil moisture prediction from the weather stations (Figure 1.3). Crops with dense, closed leaf canopies and at their peak vegetative growth stage showed higher regression coefficients compared to thin, open crop leaf canopies. This is consistent with studies showing that cropping system and crop growth stage influence soil water storage (Daigh et al., 2014) and impacts soil hydrology (Steele et al., 2012; Kravchenko et al., 2011; McIsaac et al., 2010). The type of crop and their growth stage influences small-scale soil moisture variability due to the pattern of 1) throughfall imposed by the canopy (Zheng et al., 2019), 2) shading the soil surface and affecting rate of evaporative drying (Todd et al., 1991), 3) moderating or inducing turbulence airflows and corresponding evapotranspiration rates (Katul et al., 2012), and 4) affecting soil Ksat through root distributions and their activity with extracting soil moisture for plant transpiration (Schymanski et al., 2008). The degree to which these factors affect the soil moisture depends upon plant species, density and season (Famiglietti et al., 1998; Lull and Reinhart, 1995; Reynolds 1970b. These results are in accordance with Hawley et al. (1983), Francis et al. (1986), Ozkan and Gokbulak (2017) who found significant difference in soil moisture content due to

difference in vegetation cover. For instance, row crop systems tend to have lower water storage capacities than natural or restored perennial systems (Mitchell et al., 2012; Qi et al., 2011; Brye et al., 2000) which is linked with soil moisture contents. Similarly, Gomez-Plaza et al. (2000) argued vegetated areas and vegetation cover improves soil structure and capacity of water retention into the soil compared to drier with low vegetation cover in southeastern Spain. The vegetation and land-use could have significant effect on the temporal and spatial dynamics of soil water (Qiu et al., 2001; Fu et al., 2003; Jun et al., 2010). However, Zhao et al. (2010) postulated that correlation analysis showed that soil properties were important factors controlling temporal stability of soil moisture spatial patterns for any cropping practices or vegetation cover in a semi-arid region.

Soil physical properties (bulk density, Ksat, soil texture) are well known parameters that significantly affect soil moisture. The regression analysis for soil moisture prediction showed higher r^2 values for soils with higher clay percentage as compared to sand percentage. Variation in soil texture, organic matter and macro porosity affect the water retention of soils, thereby causes the soil moisture variation (Famiglietti et al., 1998; Crave and Gascuel-Odoux, 1997; Dong and Ochsner, 2018). Similar to our findings, English et al. (2005) found sand-rich soil throughout the soil profile increases gravimetric water and soil water potential compared to clay-rich soils. Soil texture influences soil moisture through its direct effects on pore spaces governing evaporation and drainage rates, which are two main factors for controlling soil drying (Pan and Peters-Lidard, 2008; Dexter, 2004). The irrigation cycle study conducted by Li et al. (2014) showed soil water content was significantly and consistently correlated with soil texture and bulk density. Similarly, both principal component analysis and multiple linear regression identified soil texture as primary physical process controlling variability in soil water content of

agriculture field (Manns et al., 2014). Gao et al. (2011) in their study in Loess plateau, China, reported strong correlation between soil texture and surface soil moisture in gullies. Gao et al. (2011) also reported clay and silt content were both positively correlated with soil moisture during and regression values decrease with rainfall events. In the study of eleven textural classes, Vereecken et al. (2007) found that standard deviation of soil moisture peaked between 0.17 and 0.23 m^3 m^{-3} for most textual classes such as, silt loam to clay loam soils. In contrast, they found standard deviation increases with increase in soil moisture for sandy loam and loamy sand soils.

Residue cover was also correlated with soil moisture prediction. We observed high r^2 values in areas with low residue cover (<10%) and lower r^2 values in areas with high residue cover. Residue cover on the soil surface not only limit soil erosion due to water and air but also changes soil moisture spatial patterns within fields (Dabney, 1998; Daigh et al., 2019). Studies have shown that the reduction of soil evaporation due to residue cover maintains higher soil moisture contents at field level over time (Dabney, 1998; Unger and Vigil, 1998). Partial residue covers in the field contribute to a slower, but still positive effect on soil moisture recharge as compared to completely covered soils, this difference in water recharge at different residue covers affects the prediction of soil moisture at field level (Patrignani and Ochsner, 2018).

As aforementioned, many of the plant and soil characteristics in fields induce not only spatial variability, but also influence soil moisture over time. We observed that a four-day cumulative rainfall and PET had the highest non-linear regression coefficient and lowest RMSE as compared to other cumulative periods. Several studies have established similar relationships between soil moisture with the rainfall at larger spatial scales than the RRVN (Brocca et al., 2007; Entin et al., 2000; Yoo et al., 1998; Cosh et al., 2004; Ziadat and Taimeh, 2013). With the application of satellite data, Brocca et al. (2013) also established that four-day cumulative

rainfall can effectively predict soil moisture with correlation value close to 0.8, which is similar to our finding. Additionally, Brocca et al. (2007) reported that higher correlation coefficients for soil moisture as the antecedent precipitation increased, which were in accordance with Western et al. (1999) and Gomez-Plaza et al. (2001). Rainfall, as well as incoming solar radiation, are key factors affecting soil moisture at point scale measurements (Vivoni et al., 2010). Entekhabi and Rodriguez-Iturbe, (1994) and Pan et al. (2003) in their extensive studies on predicting surface soil moisture from rainfall, observed that time-weighted averages of previous cumulative rainfall over a given period resulted in high correlation coefficients with soil moisture.

Conclusion

The results shown in this study offers evidence that soil moisture can be reasonably represented by using information obtained at nearby weather stations despite large differences in soil and crop characteristics. The correlation between the soil moisture at weather stations and nearby agricultural fields is affected by crop type and their growth stages, crop residue, soil texture, and distance from the weather station. In Red River Valley, crops with thick canopy cover showed higher correlations compared to sparse crop canopies. Similar associations were observed when crop growth stages were at peak vegetative and reproductive stages. However, higher correlations were observed with lower crop residue cover of the soil surface and vice-versa. The correlation between soil moisture at weather stations and nearby fields decrease as the distance from weather stations increase. Rainfall and evapotranspiration measured at weather stations can be used to estimate soil moisture in these nearby agricultural fields. The four-day cumulative rainfall and PET showed higher correlations with field soil moisture as compared to other durations. This shows that rainfall and precipitation data can be effectively used in the prediction on soil moisture in the nearby fields despite discrepancies in soil and crop

characteristics. This study showed promising results on estimation of soil moisture on agricultural fields using nearby weather station data when considering key field variables. However, the level of effect of each of the variables on the soil moisture prediction using soil moisture of weather station needed further exploration. The use of different multivariate or machine learning algorithms to model and evaluate the influence of variables needs further exploration.

References

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., Dehnavi, S., 2017. Temperature-Vegetation-soil Moisture Dryness Index (TVMDI). Remote Sens. Environ. 197, 1-14. http://dx.doi.org/10.1016/j.rse.2017.05.026

Babaeian, E., Sadeghi, M., Jones, S.B., Montzka, C., Vereecken, H., Tuller, M., 2019. Ground, proximal, and satellite remote sensing of soil moisture. Rev. Geophys. 57(2), 530-616.

Bardossy, A., Lehmann, W., 1998. Spatial distribution of soil moisture in a small catchment. Part 1: geostatistical analysis. J. Hydrol. 206(1-2), 1-15.

Brocca L, Morbidelli, R., Melone, F., Moramarco, T., 2007. Soil moisture spatial variability in experimental areas of Central Italy. J. Hydrol. 333, 356–373. https://doi.org/10.1016/j.jhydrol.2006.09.004

Brocca, L., Moramarco, T., Melone, F., Wagner, W., 2013. A new method for rainfall estimation through soil moisture observations. Geophys. Res. Lett. 40(5), 853-858.

Brye, K.R., Norman, J.M., Bundy, L.G., Gower S.T., 2000. Water-budget evaluation of prairie and maize ecosystems. Soil Sci. Soc. Am. J. 64, 715–724. doi:10.2136/sssaj2000.642715x

Cosh, M.H., Stedinger, J.R., Brutsaert, W., 2004. Variability of surface soil moisture at the watershed scale. Water Resour. Res. 40, W12513.

Crave, A., Gascuel-Odoux, C., 1997. The influence of topography on time and space distribution of soil surface water content. Hydrol. Process. 11(2), 203-210.

Cunningham, G.M., Walker, P.H., Green, D.R., 1978. Revegetating the Cobar country – a ten-year study. J. Soil Cons. Series of New South Wales 34, 139–144.

Dabney, S.M., 1998. Cover crop impacts on watershed hydrology. J. Soil Water Conserv. 53(3), 207–213.

Daigh, A.L.M., DeJong-Hughes, J., Gatchell, D.H., Derby N.E., Alghamdi R., Leitner Z.R., Wick, A., Acharya, U., 2019. Crop and soil responses to on-farm conservation tillage practices in the Upper Midwest. Agric. Environ. Lett. 4(1), 1-5 doi:10.2134/ael2019.03.0012

Daigh, A.L.M., Ghosh, U., DeJong-Hughes, J., Horton, R., 2018. Spatial response of near-surface soil water contents to newly imposed soil management. Agric. Environ. Lett. 3:180032.

Daigh, A.L.M., Zhou, X., Helmers, M.J., Pederson, C.H., Ewing, R., Horton, R., 2014. Subsurface drainage flow and soil water dynamics of reconstructed prairies and corn rotations for biofuel production. Vadose Zone J., 13(4), 1-11.

Dalton, M., Andrews, P., Buss, P., Barrett, B., 2011. The use of the Evapotranspiration Stress Index (ETSI) to guide irrigation management in young olives. Acta Hortic. 924, 31-39 https://doi.org/10.17660/ActaHortic.2011.924.2

Dexter, A.R., 2004. Soil physical quality: Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth. Geoderma. 120(3-4), 201-214.

Dong, J., Ochsner, T.E., 2018. Soil texture often exerts a stronger influence than precipitation on mesoscale soil moisture patterns. Water Resour. Res. 54(3), 2199-2211.

Dunin, F.X., Reyenga, W., 1978. Evaporation from a Themeda grassland: I. Controls imposed on the process in a sub-humid environment. J. Appl. Ecol. 15, 317–325.

English, N.B., Weltzin, J.F., Fravolini, A., Thomas, L. and Williams, D.G., 2005. The influence of soil texture and vegetation on soil moisture under rainout shelters in a semi-desert grassland. J. Arid Environ. 63(1), 324-343.

Entekhabi, D., Rodriguez-Iturbe I., 1994. Analytical framework for the characterization of the space-time variability of soil moisture. Adv. Water Resour. 17, 35-45.

Entin, J.K., Robock, A., Vinnikov, K.Y., Hollinger, S.E., Liu, S., Namkai, A., 2000. Temporal and spatial scales of observed soil moisture variations in the extratropics. J. Geophys. Res. 105, 865–877.

Famiglietti, J.S., Rudnicki, J.W., Rodell, M., 1998. Variability in surface soil moisture content along a hillslope transect: Rattlesnake Hill, Texas. J. Hydrol. 210, 259–281.

Fernandez-Illescas, C.P., Porporato, A., Laio, F., Rodriguez-Iturbe, I., 2001. The ecohydrological role of soil texture in a water-limited ecosystem. Water Resour. Res., 37(12), 2863–2872.

Francis, C.F., Thornes, J.B., Romero Diaz, A., Lopez Bermudez, F., Fisher, G.C., 1986. Topographic control of soil moisture, vegetation cover and land degradation in a moisture stressed Mediterranean environment, Catena, 13, 211–225.

Fu, B., Wang, J., Chen, L., Qiu, Y., 2003. The effects of land use on soil moisture variation in the Danangou catchment of the Loess Plateau, China. Catena 54, 197–213.

Gao, X., Wu, P., Zhao, X., Shi, Y., Wang, J., Zhang, B., 2011. Soil moisture variability along transects over a well-developed gully in the Loess Plateau, China. Catena, 87(3), 357-367.

Gee, G.W., Bauder, J.W., 1986. Particle Size Analysis. In: Methods of Soil Analysis, Part A. Klute (ed.). 2 Ed., Vol. 9. Am. Soc. Agron., Madison, WI, 383-411.

Gomez-Plaza, A., Alvarez-Rogel, J., Albaladejo, J., Castillo, V., 2000. Spatial patterns and temporal stability of soil moisture across a range of scales in a semi-arid environment. Hydrol. Process. 14(7), 1261-1277.

Gomez-Plaza, A., Martinez-Mena, M., Albaladejo, J., Castillo, V., 2001. Factors regulating spatial distribution of soil water content in small semiarid catchments. J. Hydrol. 253, 211–226.

Gwak, Y., Kim, S., 2017. Factors affecting soil moisture spatial variability for a humid forest hillslope. Hydrol. Process., 31(2), 431-445.

Hamman, B., Egil, D.B., Koning, G., 2002. Seed vigor, soilborne pathogens, pre-emergent growth, and soybean seeding emergence. Crop Sci. 42, 451-457. https://doi.org/10.2135/cropsci2002.0451

Hawley, M.E., Jackson, T.J., McCuen, R.H., 1983. Surface soil moisture variation on small agricultural watersheds. J. Hydrol. 62, 179–200.

Helms, T.C., Deckard, E., Goos, R.J., Enz, J.W., 1996. Soybean seedling emergence influenced by days of soil water stress and soil temperature. Agron. J. 88, 657-661.

Hu, Z., Islam, S., Cheng, Y., 1997. Statistical characterization of remotely sensed soil moisture images. Remote Sens. Environ. 61, 310–318.

Ivanov, V.Y., Fatichi, S., Jenerette, G.D., Espeleta, J.F., Troch, P.A., Huxman, T.E., 2010. Hysteresis of soil moisture spatial heterogeneity and the “homogenizing” effect of vegetation. Water Resour. Res., 41, W09521, doi:10.1029/2009WR008611.

Jun, F., Mingan, S., QuanJiu, W., Jones, S.B., Reichardt, K., Xiangrong, C., Xiaoli, F., 2010. Toward sustainable soil and water resources use in China’s highly erodible semi-arid loess plateau. Geoderma 155, 93–100.

Kansas Mesonet, 2020. https://mesonet.k-state.edu/.

Katul, G.G., Oren, R., Manzoni, S., Higgins, C., Parlange, M.B., 2012. Evapotranspiration: a process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. Rev. Geophys. 50(3), 1-25.

Kravchenko, A.N., Wang, A.N.W., Smucker, A.J.M., Rivers, M.L., 2011. Long-term differences in tillage and land use affect intra-aggregate pore heterogeneity. Soil Sci. Soc. Am. J. 75, 1658–1666. doi:10.2136/sssaj2011.0096

LaGuardia, G., Niemeyer, S., 2008. On the comparison between the LISFLOOD modelled and the ERS/SCAT derived soil moisture estimates. Hydrol. Earth Syst. Sci., 12(6), 1339-1351.

Li, T., Hao, X.M., Kang, S.Z., 2014. Spatiotemporal variability of soil moisture as affected by soil properties during irrigation cycles. Soil Sci. Soc. Am. J. 78(2), 598–608. https://doi.org/10.2136/sssaj2013.07.0269

Lull, H.W., Reinhart, K.G., 1955. Soil moisture measurement. U.S.D.A. Southern For. Exp. Sta., New Orleans, LA., Occas. Paper No. 140.

Manns, H. R., Berg, A. A., Bullock, P. R., McNairn, H. 2014. Impact of soil surface characteristics on soil water content variability in agricultural fields. Hydrological Processes, 28(14), 4340–4351. https://doi.org/10.1002/hyp.10216

McIsaac, G.F., David, M.B., Mitchell, C.A., 2010. Miscanthus and switchgrass production in central Illinois: Impacts on hydrology and inorganic nitrogen leaching. J. Environ. Qual. 39, 1790–1799. doi:10.2134/jeq2009.0497

McMillan, H.K., Srinivasan, M.S., 2015. Characteristics and controls of variability in soil moisture and groundwater in a headwater catchment. Hydrol. Earth Syst. Sci. 19, 1767‑1786.

Mitchell, J.P., Singh, P.N., Wallender, W.W., Munk, D.S., Wroble, J.P., Horwath, W.R., 2012. No-till and high-residue practices reduce soil water evaporation. Calif. Agric. 66(2), 55-61. doi:10.3733/ca.v066n02p55

NDAWN, 2020. North Dakota Agricultural Weather Station. https://ndawn.ndsu.nodak.edu/ (data retrieved on 5/8/2020).

NOAA/NCEI, 2020. National Oceanic and Atmospheric Administration/National Centers for Environmental Information. https://www.ncdc.noaa.gov/ (Data retrieved on 7/15/2020)

NWS, 2020. National Weather Services. https://www.weather.gov/ (data retrieved on 5/8/2020)

Oklahoma Mesonet, 2020. http://www.mesonet.org/index.php.

Ozkan, U., Gökbulak, F., 2017. Effect of vegetation change from forest to herbaceous vegetation cover on soil moisture and temperature regimes and soil water chemistry. Catena, 149, 158-166.

Pan, F., Peters-Lidard, C.D., 2008. On the relationship between mean and variance of soil moisture fields. J. Am. Water Resourc. Assoc. 41(1), 235–242, doi:10.1111/j.1752-1688.2007.00150.x.

Pan, F., Peters-Lidard, C.D., Sale, M.J., 2003. An analytical method for predicting surface soil moisture from rainfall observations. Water Resour. Res. 39(11), 1-12.

Patrignani, A., Ochsner, T.E., 2018. Modeling transient soil moisture dichotomies in landscapes with intermixed land covers. J. Hydrol. 566, 783-794.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A. Ma. P. Scis., 193(1032), 120-145.

Qi, Z., Helmers, M.J., Christianson, R.D., Pederson C.H., 2011. Nitrate-nitrogen losses through subsurface drainage under various agricultural land covers. J. Environ. Qual. 40, 1578-1585. doi:10.2134/jeq2011.0151

Qiu, Y., Fu, B., Wang, J., Chen, L., 2001. Soil moisture variation in relation to topography and land use in a hillslope catchment of the Loess Plateau, China. J. Hydrol. 240, 243–263.

Reynolds, S.G., 1970. The gravimetric method of soil moisture determination, I: A study of equipment, and methodological problems. J. Hydrol. 11(3), 258-273.

Reynolds, S.G., 1970b. The gravimetric method of soil moisture determination, II: Typical required sample sizes and methods of reducing variability, J. Hydrol. 11, 274–287.

Richards, L.A., 1948. Porous plate apparatus for measuring moisture retention and transmission by soil. Soil Sci. 66, 105-110. doi: 10.1097/00010694-194808000-00003

Rosenbaum, U., Bogena, H.R., Herbst, M., Huisman, J.A., Peterson, T.J., Weuthen, A., Western, A.W., Vereecken, H., 2012. Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale. Water Resour. Res., 48(10), 1-22.

SAS, 2017. SAS Institute version 9.4. Foundation for Microsoft Windows. SAS Inst., Cary, NC.

Schaap, M.G., Leij, F.J., vanGenuchten, M.Th., 2001. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol., 251, 163-176.

Schymanski, S.J., Sivapalan, M., Roderick, M.L., Beringer, J., Hutley, L.B., 2008. An optimality-based model of the coupled soil moisture and root dynamics. Hydrol. Earth Syst. Sc. 12(3), 913-932.

Simunek, J., Sejna, M., Saito, H., Sakai, M., vanGenuchten, M.Th., 2008. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media version 4.0. Department of Environmental Sciences, University of California Riverside, California.

Steele, M.K., Coale, F.J., Hill, R.L., 2012. Winter annual cover crop impacts on no-till soil physical properties and organic matter. Soil Sci. Soc. Am. J. 76, 2164-2173. doi:10.2136/sssaj2012.0008

Thompson, S.E., Harman, C.J., Heine, P., Katul, G.G., 2010. Vegetation-infiltration relationships across climatic and soil type gradients. J. Geophys. Res-Biogeo. 115, 1-12.

Todd, R.W., Klocke, N.L., Hergert, G.W., Parkhurst, A.M., 1991. Evaporation from soil influenced by crop shading, crop residue, and wetting regime. T. ASAE. 34(2), 461-0466.

Unger, P.W., Vigil, M.F., 1998. Cover crop effects on soil water relationships. J. Soil Water Conserv. 53, 200–207.

USDA. 2020. United States Department of Agriculture, International Production Assessment Division. Metadata for crops at different growth stage.

https://ipad.fas.usda.gov/cropexplorer/description.aspx?legendid=312 (data retrieved on 5/8/2020)

Vereecken, H., Kamai, T., Harter, T., Kasteel, R., Hopmans, J., Vanderborght, J., 2007. Explaining soil moisture variability as a function of mean soil moisture, a stochastic unsaturated flow perspective. Geophys. Res. Lett., 34(22), 1-6, doi:10.1029/2007GL031813.

Verstraeten, W.W., Verousraete, F., Feyen, J., 2007. Assessment of evapotranspiration and soil moisture content across different scales of observation. Sensors. 8, 70-117.

Vivoni, E.R., Rodríguez, J.C., Watts. C.J., 2010. On the spatiotemporal variability of soil moisture and evapotranspiration in a mountainous basin within the North American monsoon region. Water Resour. Res. 46, W2509, doi:10.1029/2009WR008240.

Western, A.W., Grayson, R.B., Blo¨schl, G., Willgoose, G.R., McMahon, T.A., 1999. Observed spatial organization of soil moisture and its relation to terrain indices. Water Resour. Res. 35 (3), 797–810.

Western, A.W., Grayson, R.B., Blöschl, G., 2002. Scaling of soil moisture: A hydrologic perspective. Annu. Rev. Earth Planet. Sci., 30, 149–180. doi:10.1146/annurev.earth.30.091201.140434.

Yoo, C., Valdes, J.B., North, G.R., 1998. Evaluation of the impact of rainfall on soil moisture variability. Adv. Water Resour. 21, 375–384.

Zhao, Y., Peth, S., Wang, X.Y., Lin, H., Horn, R., 2010. Controls of surface soil moisture spatial patterns and their temporal stability in a semi-arid steppe. Hydrol. Process. 24, 2507-2519.

Zheng, J., Fan, J., Zhang, F., Yan, S., Wu, Y., Lu, J., Guo, J., Cheng, M., Pei, Y., 2019. Throughfall and stemflow heterogeneity under the maize canopy and its effect on soil water distribution at the row scale. Sci. Total Environ. 660, 1367-1382.

Ziadat, F.M., Taimeh, A.Y., 2013. Effect of rainfall intensity, slope, land use and antecedent soil moisture on soil erosion in an arid Environment. Land Degrad. Dev. 24(6), 582-590.

CHAPTER II. MACHINE LEARNING FOR PREDICTING FIELD SOIL MOISTURE USING SOIL, CROP, AND NEARBY WEATHER STATION DATA IN RED RIVER VALLEY OF NORTH

Abstract

Soil moisture plays important role in agricultural production and hydrological cycles and its precise prediction is important for water management and logistics of on-farm operations. However, soil moisture is affected by various soil, crop and meteorological factors, and it is difficult to establish ideal mathematical models for soil moisture prediction. In this study, we investigate various machine learning techniques for predicting soil moisture in the Red River Valley of North (RRVN). Specifically, the machine learning techniques evaluated include: Classification and Regression Trees (CART), Random Forest Regression (RFR), Boosted Regression Trees (BRT), Multiple Linear Regression (MLR), Support Vector Regression (SVR) and Artificial Neural Network (ANN). The objective of this study was to determine the effectiveness of these machine learning techniques and evaluate the importance of predictor variables. Variables to predict field soil moisture included soil texture, bulk density, saturated hydraulic conductivity (Ksat), crop type and growth stage, crop residue cover, four-day cumulative rainfall and potential evapotranspiration (PET). The RFR and BRT algorithms performed the best with mean absolute errors (MAE) of < 0.040 $m^3 m^{-3}$ and root mean square errors (RMSE) of 0.045 and 0.048 $m^3 m^{-3}$, respectively. Similarly, RFR, SVR, and BRT showed high correlations (r^2 of 0.72, 0.65 and 0.67 respectively) between predicted and measured soil moisture. The MLR and ANN had the poorest performance (r^2=0.52, RMSE=0.059 and r^2=0.53, RMSE=0.085, respectively). The CART, RFR, and BRT models showed soil moisture at nearby weather stations had the highest relative influence for moisture prediction, followed by the four-

day cumulative rainfall and PET, and subsequently followed by bulk density and Ksat. The RFR, SVR, and BRT algorithms showed promising results in soil moisture prediction using soil, crop, and weather station variables in RRVN. Soil moisture, four-day cumulative rainfall and PET, bulk density and Ksat can be effectively used in soil moisture prediction of nearby field. Therefore, machine learning models, provided with few weather stations, crop and soil data can effectively predict soil moisture of nearby agricultural fields.

Key words: Machine learning algorithms, Vertisols, Mollisols, soil moisture prediction, weather station.

Introduction

Soil moisture has a strong influence on the distribution of water between various components of hydrological cycle in agricultural field. It helps in understanding the hydrology and climate that have high spatial and temporal variability. Precise measurement and/or prediction of soil moisture provides insights in expected infiltration and runoff generation during rainfall event and management of the water for agricultural purpose (Gill et al., 2006). In agricultural field, soil moisture affects key farm activities from crop selection to timing of tilling, planting, fertilizer application and harvesting due to infiltration, evaporation, runoff, heat and gas fluxes (Amani et al., 2017; Hamman et al., 2002). Soil moisture prediction across large spatial scales is difficult due to the heterogeneity in soil texture, crop type, crop residue cover. Point measurement that includes gravimetric method, in-situ electromagnetic sensors are accurate but have limited spatial extent and need a lot of time and labor (Laguardia and Niemeyer, 2008).

In practice, farmers typically rely on heuristic approaches with weather station data (i.e., rainfall, evapotranspiration, temperature) to predict (or extrapolate) the conditions in their crop fields. More accurate and optimal computational approaches need to explicitly consider various

factors such as crop type, soil texture, saturated hydraulic conductivity, and residue content affecting the soil moisture in these crop fields.

Soil moisture is predicted using information collected from nearby weather station and variables from soil and crop using one of three empirical, regression and machine learning methods (Cai et al., 2019). These methods include forecasting models such as empirical formula (Sanuade et al., 2020), water balance approach, dynamic soil water models (Zhou, 2007), time series models (Zhang et al., 2008), and neural network models (Huang et al., 2010). Traditional models include statistical regression techniques to develop geospatial functions from in-situ measurements of target and predictor variables. The advantage of traditional models is that they are typically fast to derive and do not require many inputs (Ali et al., 2015). While the disadvantage of traditional models is the need for an abundance of ground measurements that could be time consuming and expensive. Moreover, traditional modeling approach follow strict statistical assumptions and data requirements that frequently utilize linear and additive modeling approach that are not consistent with natural processes (Clapcott et al., 2013).

Recently, the use of machine learning techniques has gained attention because they can overcome some of the limitations of traditional and physics-based models. Ali et al. (2015) suggested machine learning models provide the benefit to understand and estimate complex non-linear mapping of the data distributed without any presupposition. In addition to that, this also helps to combine various sources that are poorly defined and have unknown probability functions. However, machine learning algorithms provide no information on how they have established relationships between different variables and only perform better with the large number of data sets used for training. The machine learning technique is rapidly growing in predictive modeling to identify complex data structures, which are often non-linear, and

generating accurate predictive models (Olden et al., 2008; Naghibi et al., 2016). Machine learning models have greater power for resolving and establishing complex relations (nonlinear, nonmonotonic, multimodal relationships common with landscape and ecological applications) as they are not restricted to traditional assumptions about data characteristics. There are numerous machine learning algorithms such as, classification and regional tress (CART), random forest regression (RFR) (Liaw and Wiener, 2002), support vector regression (SVR) (Zaman et al., 2012; Zaman and Mckee, 2014), multiple linear regression modeling, boosted regression tree (BRT), artificial neural networks (ANN) (Hassan-Esfahani et al., 2015), etc. that are used for predictions. In the hydrology domain, neural networks (Qiao et al., 2014), vector machines (Kashif Gill et al., 2007), and polynomial regression (Gorthi and Dou, 2011) has been used in soil moisture prediction using historical soil moisture datasets.

For example, Matei et al. (2017) used different machine learning models (SVR, NN, LR, RFR, etc.) for real time soil moisture prediction in Transylvania Depression of Romania. They used data (soil temperature, air temperature, precipitation) from a nearby weather station and used crop and soil information nearby station. Machine learning-based model (i.e., an RFR) achieve better performance when compared with the physics-based Richards equation model in predicting soil matric potential in the root zone (Gumiere et al., 2020). Yoon et al. (2011) used ANN to model the water table dynamics of various agricultural systems. Random Forest model was found to be superior to ANN model when predicting lake water levels with fewer parameters and training time (Li et al., 2016). Alternatively, Gill et al. (2006) used SVM to predict soil moisture using meteorological data, field data and crop data. The SVM employs structural risk minimization instead of the traditional risk minimization that formulates quadratic optimization to ensure a global optimum. The resulting SVM model is sparse and not affected by the

dimensionality. Support Vector regression is less prone to overfitting the regression function because it uses generalize error bound (ε)-insensitive loss function and structural risk optimization (Vapnik, 2000).

Machine learning algorithms

Classification and regression trees (CART)

CART is a tree-based regression model and rule-based procedure that creates a binary tree using binary recursive partitioning (BRP) to yield the maximum reduction in the variability of the response variable (Stewart, 1996). The BRP is a nonparametric nonlinear technique that splits the data into subsets based on available independent factors, which means it divides nodes into yes/no answers as predictor values. Regression trees are generated for continuous data and classification trees for categorical data (Samadi et al., 2014).

Suppose input variables are x_1, x_2, …. x_n and output variable is y for training dataset in the space D with n input variables and m input samples. Let $D = \{(x_{11}, x_{12}, \ldots.x_{1n}, y_1), (x_{21}, x_{22}, \ldots.x_{2n}, y_1), (x_{m1}, x_{m2}, \ldots.x_{mn}, y_m)\}$. The CART model splits D into certain number of subspaces using a BRP. Each recursive process attempts to select several splitting variables and splitting points from the current space S (parent node) to divide the space into two inhomogeneous subspaces S_1 and S_2. Every subspace has an estimated value $\hat{y}$ determined by fitting using least square method; the optimal splitting variable j and splitting point s are finally selected to ensure that the binary division has the minimum residual variance (Breiman et al., 1984).

Random forest regression (RFR)

Random forest regression was developed by Breiman (2001) and are relatively simple to train, tune, and apply. This technique is developed as average over the many individual decision tree-based models that are built on the bootstrapped training sample. Each training sample

considers a small group of predictor variables at every split that maintains decorrelation among the variables and splits (James et al., 2013). The RFR improves predictive accuracy by generating large numbers of decision trees that classify a case using each tree in the new forest and deciding a final predicted outcome by combining the results across all the trees. Each tree is built using a deterministic algorithm by selecting a random set of variables and a random sample from the training dataset (i.e., the calibration data set). The RFR uses three parameters, 1) the number of regression trees grown based on the bootstrap sample of the observation (e.g., hundreds or thousands of trees), 2) the number of different predictors tested at each node (e.g., one third of the total number of the variables) and 3) the minimal size of the terminal nodes of the trees (e.g., one).

Boosted regression trees (BRT)

Friedman (2002) defined BRT as a decision tree model that is improved by the gradient boosting algorithm, which constructs an additive regression model that fits in chronological order based on simple base learner function to current pseudo-residuals at each iteration. The pseudo-residuals are defined as the slope of the loss function that is being minimized. Due to the use of pseudo-residual, simple base learner function and iteration at each level, this model has performed better compared to other machine learning models (Elith et al., 2008; Natekin and Knoll, 2013). Due to its ability to perform better in complicated data, BRT models are popular and attractive among data scientists (Friedman, 2001), whereas, the data used for training sets are compiled from different sources makes it susceptible to some kind of inconsistencies (Breiman, 2001; James et al., 2013).

The BRT helps in partitioning influences of the independent (predictor) variables on the dependent variable (soil moisture for this study). This combination of regression trees with the

boosting algorithm has been used by ecologists to explore the relationship between ecological processes and predictors. The BRT handles predictor variables with different data types, distributions, and completeness (i.e., level of missing values) (Zhang et al., 2015). The fitting of the BRT model is controlled by different factors such as: the learning rate that determines the contribution of each tree to the growing model, the tree complexity that controls the level of interactions in BRT, the bagging fractions that sets the proportion of observations used in selecting variables, and the cross validation that specifies the number of times to randomly divide the data for model fitting and validation (De'Ath, 2017).

Multiple linear regression (MLR)

Multiple linear regression is an extension of simple linear regression used to predict an outcome variable based on multiple distinct predictor variables. With 'n' number of predictor variables (x), the predictions of y are expressed by the following equation.

$$y = b_0 + b_1 x_1 + b_2 x_2 \ldots\ldots b_n x_n \quad (2.1)$$

The b values are called the regression weights (beta coefficients). The measures of association between the predictor variable and the outcome. 'b_i' can be interpreted as the average effect on y of a unit increase in x_i, holding all other predictors fixed.

Support vector regression (SVR)

Support vector regression uses a support vector machine (SVM) to solve regression problems (Cortes and Vapnik, 1995; Drucker et al., 1997). SVM learning simplifies a maximal margin classifier to map the input variables into a high dimensional space using fixed mapping kernel function. To overcome local minima problems created due to use of few parameters while tuning the training dataset, SVR uses radial basis kernel function by constructing the hyperplanes that can be used for regression. Yang et al. (2009) found radial basis functions powerful because they are simple and reliable and deal with complex dimensional space, margin separation factors.

The SVR involves use of a sub-set of data points that are based on a predefined error margin to fit a regression model between dependent variable and explanatory variables. Those sub-set of data points are called support vectors. Let us suppose, there are given samples $X = \{\vec{x}_1, \vec{x}_2,\ldots,\vec{x}_l\}$ and corresponding target value $Y = \{y_1, y_2,\ldots,y_l\}$, $y_i \in R$. The goal of SVR is to find the function $f(x)$ that has at most standard deviation from the obtained target y_i for all training data and meanwhile as flat as possible.

Artificial neural network (ANN)

Artificial neural networks are a powerful computing tool constructed through many simple interconnected elements called neurons with unique capability of recognizing underlying relationship with input and output events. An ANN consist of input, hidden and output layers arranged in a discrete manner (Priddy and Keller, 2005). A collection of neurons arranged in the dimensional array is called a layer, where each layer includes one or more individual nodes. The number of input variables necessary for predicting the desired output variables determines the number of input nodes. The complexity of modeling is dependent upon the optimum number of hidden nodes and hidden layers (i.e., the greater the number of hidden layer results in a larger more complex model) (Grimes et al., 2003).

These ANN models mimic the human-learning ability by learning from a training data set. They handle robust to noisy data and create a powerful tool to approximate multivariate non-linear relations among the variables (Twarakavi et al., 2006). ANN is powerful tool that can approximate all types of non-linear mapping. These models have been used for input-output correlations of non-linear processes in water resources and hydrology (Ahmad and Simonovic, 2005).

Objectives of the study

The goal of this study was to determine the performance of the aforementioned machine learning models for predicting soil moisture of crop fields throughout the Red River Valley of the North by using weather station observations and field characteristics of nearby areas under crop management. The objective of this study was to (i) evaluate the effectiveness of different machine learning tools in soil moisture prediction, (ii) find out the important predictor variables affecting field soil moisture content using machine learning tools.

Methods

Study site and weather station

This study was conducted along the Red River Valley of the North (RRVN) in North Dakota and Minnesota. The RRVN is a glaciolacustrine lakebed formed by ancient Lake Agassiz that existed for more than 4,000 years. The topography is minimal (1 meter per 5 kilometer) and Mollisols and Vertisols are the dominant soil orders with soil texture ranges from clay to loamy sand. The parent material for RRVN is poorly drained and consist of gray, slickensided, flat clays of Brenna/Argusville formation that are overlain by the tan-buff, laminated silty clays of the Sherack formation. The major corps cultivated in this area includes corn, soybean, wheat, sugarbeet, barley, canola and potato. The annual mean temperature is 4 °C typically varies from -16 °C to 29 °C and rarely below -27 °C or above 32 °C, whereas the 30 year mean annual rainfall is 60 cm and snowfall of 125 cm (NOAA/NCEI, 2020). Summers are long and warm whereas, winters are frigid, snowy, windy, and partly cloudy year-round.

There are 117 Weather stations under North Dakota Agricultural Weather Network (NDAWN) in North Dakota (83), Minnesota (28) and Montana (6), which reports 32 weather parameters (e.g., air temperature, rainfall, wind direction, soil moisture). Our study area covers a

total of 25 weather stations, where 15 stations are located across 8 counties of North Dakota and 10 stations are located across 7 counties in Minnesota (Figure 2.1). Weather station data and measurements in nearby agricultural fields of study area were collected during the cropping season from June to September, 2019. Soil moisture was measured around the weather station and nearby crop field in 16 days interval to coincide with satellite imagery pass dates. The distance between crop field and weather station was measured in meters. The distance measure was classified into five different classes (0-100 m, 100-200 m, 200-400 m, 400-800 m, 800-1200 m and 1200-2000m). The crop field under study were within the range of 2000 m of the weather station for entire study area of the RRVN.

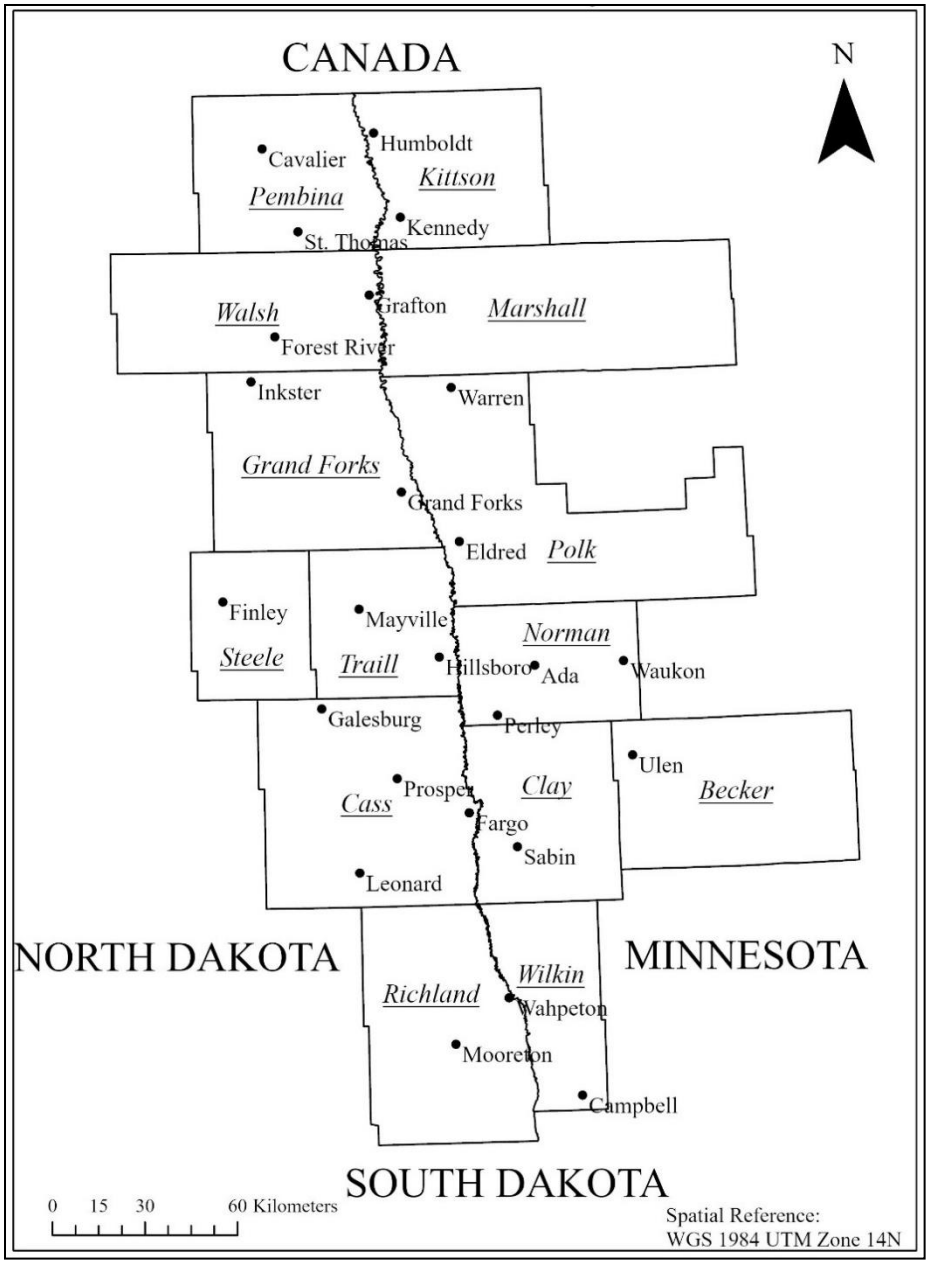

Figure 2.1. Map showing counties of North Dakota and Minnesota and weather stations under study area around Red River Valley. Black dots in map represents weather stations and italic with underline word represents counties

Soil moisture measurement

Soil moisture from each field and weather station was measured using gravimetric method. Soil samples were collected from field using Uhland cores. Composite soil sample was collected from three different location of individual field using sampling core of dimensions 6

cm × 8 cm at 0 to 6 cm depth and GPS coordinates were recorded. The weight of wet soil samples collected from field were taken and oven dried at 105°C for 48 hours. Gravimetric water content was determined using dry weight of soil and amount of water loss during the drying. Volumetric water content (VWC) of soil was calculated by multiplying gravimetric water content ($m^3\ m^{-3}$) with bulk density ($g\ cm^{-3}$) (Reynolds, 1970).

Crop types in the study area

This study covers all types of crops grown in this area such as soybean (24 plots), wheat (18 plots), corn (16 plots), sugar beet (6 plots), dry beans (5 plots), oats (2 plots), barley (plots 1), potato (plots 1), canola (plots 1) and alfalfa (plots 1). Soil samples for bulk density and moisture content were collected after the germination of the crops starting from first week of June, 2019. Soil samples were taken in 16 days intervals and growth stages for each crop were recorded using the standards developed by United States Department of Agriculture (USDA, 2020).

Residue cover, soil texture and saturated hydraulic conductivity

Antecedent characteristics such as residue cover, soil texture and saturated hydraulic conductivity (Ksat) was determined for each location from where soil samples were collected. Residue cover was determined using the rope method along eight transects per sample site (i.e., residue presence at 100 points along 15 m oriented 45° to plant rows) (Daigh et al. 2019). The crop residue was then pooled and classified as percentage in three different categories (<10%, 20-30% and 50-60% residue cover) for analysis. Soil texture for this experiment was determined for each soil sample using pipette method described by Gee and Bauder (1986). Ksat ($inch\ hr^{-1}$) was estimated using Rosetta neural network pedotransfer function in Hydrus-1D that uses input data of sand, silt, clay percent, bulk density ($g\ cm^{-3}$), water contents at 33 and 1500 kPa suctions

(cm^3 cm^{-3}) (Schaap et al., 2001; Simunek et al., 2008). Pressure plate apparatus were used to determine water contents at 33 and 1500 kPa suctions (e.g. Richards, 1948).

Rainfall and potential evapotranspiration

Rainfall and potential evapotranspiration as recorded by the weather station were downloaded from the North Dakota Agricultural weather network (NDAWN) (https://ndawn.ndsu.nodak.edu/) for each weather station. Rainfall was measured hourly at a one-meter height above soil surface using TE525 tipping bucket rain gauges (Texas Electronics TR-525I, Dallas Texas) at every NDAWN weather station. Each bucket tip measures 0.254 millimeters of rainfall. Potential evapotranspiration (PET) is the estimate of the maximum daily crop water loss when water is readily available. It is calculated using Penman equation (Penman, 1948) that use soil radiation, dew point temperature, and wind speed and air temperature. The four-day cumulative rainfall and PET was calculated by adding preceding 4 days values of rainfall and PET in mm.

Machine learning procedures

The machine learning procedures were done using the R environment software (R Development Core Team 2020). The *caTools* R package was used to handle training and testing the dataset and *Metrics* package was used to calculate RMSE and MAE for all models.

Karatzolou et al., (2004) suggested the use of *kernlab* R-package along with eps-regression SVM type, radial kernel, cost value of 1, gamma value of 0.04167 and epsilon value of 0.1 to execute SVR function for analysis. Liaw and Wiener (2015) have used *randomForest* R-package to implement random forest regression model and used values for *ntree, mtry* and *nodesize* as 1000, 4 and 1, respectively.

BRT algorithm was implemented using *gbm* R package and CART algorithm using *rpart* R package. Similarly, *neuralnet* package (Gunther and Fritsch, 2010) was used for ANN algorithm, which depends on two other packages *grid* and *MASS* (Venables and Ripley, 2002). The *metrics* R package is used in supervised machine learnings and it implements metrics for regression, classification, and information retrieval problems. The *iml* R-package is used for predictor variable importance and accumulated local effects plots.

The entire set of data are divided in 70-30%, where 70% of data are used as training set and 30% of data are used for testing (i.e., validation). The testing set was used to evaluate final trained models. All the training sets have samples from all the seven sampling dates throughout the study period. Twelve predictor variables were used as input variables viz., station soil moisture, crop type, crop residue content, four-day cumulative rainfall and PET, station bulk density and Ksat, field bulk density and Ksat, sand, silt and clay percent to predict field soil moisture (dependent variable).

Statistical analysis

Model performance

The model was developed based on the training dataset and the performance of the model is evaluated based on the testing dataset also known as validation. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (r^2) are used as tool to measure the performance of different models and are determined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N}}, \tag{2.3}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}, \tag{2.4}$$

Where N, y, $\hat{y}$, $\bar{y}$ denotes number of observations, measured value, predicted value and mean of measured value, respectively. Scatter plot and box plots were used to show relationship between the observed and predict soil moisture for different machine learning models.

Soil moisture (VWC) of crop field was set as the dependent variable, whereas predictor variables were VWC of weather station, bulk density and Ksat of crop field and weather station, crop type, distance from crop field to weather station, sand, silt and clay percent of crop field, residue cover, four-day cumulative rainfall, and PET.

Variable importance

Each predictor variable has an impact on the model generated by the machine learning algorithm, the statistical significance can be measured using the variable importance. RFR algorithms calculate variable importance internally in the form of increase on the RMSE, whereas, BRT determines variable importance as percentage. On the other hand, the CART model derives the variable importance as the relative contribution of the predictor variable to field soil moisture.

Loss of mean absolute error (MAE) can characterize the influence of the predictor variable in the generated model and showed a level of effect if not considered on the prediction model. This loss of MAE was implemented using the *iml* R-package and showed 5% and 95% quartile of the loss of MAE due to the particular model.

Effect of predictor variables

Apley (2016) and Greenwell (2017) have proposed the concept of accumulated local effects (ALE) plots to establish relationship between the predictor variables and generated output. ALE plots are used to study the relationship between the outcome of machine learning models and predictor variables. Machine learning algorithms can use a lot of variables in

prediction models but few variables have huge impacts when compared to others. In this study, the top four predictor variables were selected based on the variable importance and ALE plots were created using *iml* R-package. ALE plots construct unbiased plots even when the variables under study are correlated (Apley, 2016). ALE values in the plot showed the effect of a variable on the outcome at certain values when compared to the average prediction and center values indicate the mean effect as zero. Molnar (2019) gave an example on ALE estimate of -2 when variable interest has value of 3, then it can be understood that the prediction is lowered by 2 compared to average prediction.

Result and Discussion

Model performance

Machine learning algorithms tested for performance based on the MAE, RMSE and r2 are presented in Table 2.1. The best performance was observed under RFR and BRT models and have values of MAE less than 4%. The RMSE of the predicted soil moisture using five difference machine learning algorithm ranges from 0.045 to 0.085 $m^3\ m^{-3}$. The RFR model outperformed other models based on the lowest RMSE value of 0.045 $m^3\ m^{-3}$ and higher r^2 value of 0.72. After RFR model, SVR and BRT performed well based on the RMSE (0.050 $m^3\ m^{-3}$, 0.048 $m^3\ m^{-3}$), MAE (0.039 $m^3\ m^{-3}$, 0.037 $m^3\ m^{-3}$) and r^2 (0.65, 0.67) values, respectively.

The soil moisture estimates from difference models for testing phase are shown in Figure 2.2. The RFR, BRT and SVR model performed reasonably well in capturing the soil moisture prediction in the scatter plot diagrams. The RFR model can capture the extremes (low and high values) in soil moisture content depicted by most of the sample points lying on and around the bisector line. There are few sample points which lie far away from the bisector line representing poor estimates (too high or too low). The soil moisture estimates for the BRT and SVR models

showed they can depict soil moisture prediction with slope values of 0.69 and 0.65, respectively. MLR and ANN models showed poorer performance (0.58, 0.53) in terms of spread along the bisector line for the test data.

Table 2.1. Comparison of the machine learning algorithms for soil moisture prediction using coefficient of determination (r^2), root mean squared error (RMSE) and mean absolute error (MAE). Algorithms included classification and regression trees (CART), multiple linear regression (MLR), random forest regression (RFR), support vector regression (SVR), boosted regression trees (BRT), and artificial neural networks (ANN).

Algorithms	r^2	RMSE	MAE
CART	0.57	0.056	0.045
MLR	0.52	0.059	0.046
RFR	0.72	0.045	0.034
SVR	0.65	0.050	0.039
BRT	0.67	0.048	0.037
ANN	0.53	0.085	0.068

Box plots depicting the median and percentiles (5^{th}, 25^{th}, 75^{th} and 95^{th}) of the testing data set for both the measured and predicted soil moisture is shown in Figure 2.3. The horizontal line inside the box shows the median value and box represents the 25^{th} and 75^{th} percentile (interquartile range) values whereas the whiskers extend from 5^{th} to 95^{th} percentile values. The dashed line inside the box represents the mean of the measured data for testing phase. The RFR model shows that the mean of the measured soil moisture is represented by the median of the estimated soil moisture. The RFR model performs well for estimating both low and high soil moisture values as the 5^{th} and 95^{th} percentile whisker of the measured soil moisture and predicted soil moisture satisfactorily matches. RFR followed by SVR and BRT models were able to capture the relationship between the soil moisture at each field with VWC as recorded at each pertinent weather station, rainfall, PET, crop and soil factors with lower RMSE and MAE.

The RMSE, r^2 and MAE values for the six different machine learning model showed RFR, SVR and BRT has RMSE less 0.05 $m^3\ m^{-3}$ and satisfactory r^2 values (0.65-0.67) whereas,

remaining three models had RMSE higher than 0.05 m^3 m^{-3} and r^2 lower than 0.60. This showed RFR as the best model compared to other machine learning models due to powerful averaging capacity of all the random trees generated by the model based on the number of parameters used to predict soil moisture (Matei et al., 2017). The results for scatter (Figure 2.2) and box plots (Figure 2.3) indicated that the difference in soil moisture estimates can be due to the cumulative effect of soil and crop types and change in micro-climate variables (i.e., rainfall and PET). The relatively high accuracy of RFR and BRT models is consistent with other studies that find ensemble decision-based regression models perform better than many other machine learning models (Caruana and Niculescu-Mizil, 2006); particularly in terrain and soil spatial predictions (Hengl et al., 2018; Nussbaum et al., 2018; Keskin et al., 2019; Szabo et al., 2019; Araya et al., 2020). The scatter plots and box plots showed that CART, MLR, and ANN models do not capture the extreme values in the prediction of soil moisture as well as the RFR, BRT and SVR models. Results showed that machine learning model such RFR, BRT and SVR outperformed ANN, MLR and CART models, possibly due to the carefully selected parameter optimization algorithm used to train the model. Superiority of RFR and SVR over the other models has also been reported various studies (Kalra and Ahmad, 2009; Dibike et al., 2001; Asefa et al., 2006; Gill et al., 2006; Liong and Sivapragasam, 2002; Achieng, 2019).

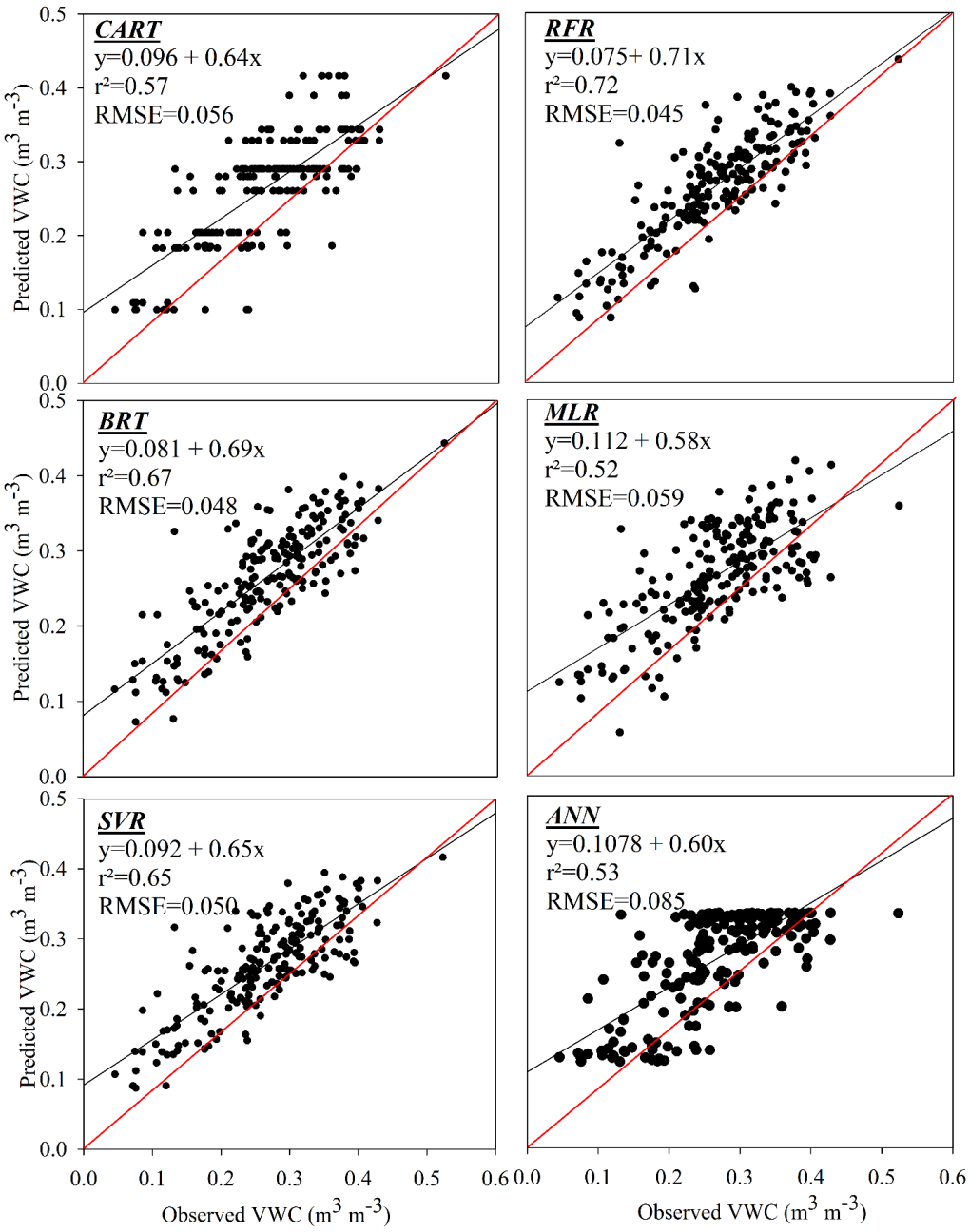

Figure 2.2. Scatter plot showing observed versus predicted volumetric water content ($m^3\ m^{-3}$) during the testing phase along with regression coefficient (r^2) and root mean square error (RMSE) for six different machine learning models. Algorithms included classification and regression trees (CART), multiple linear regression (MLR), random forest regression (RFR), support vector regression (SVR), boosted regression trees (BRT), and artificial neural networks (ANN).

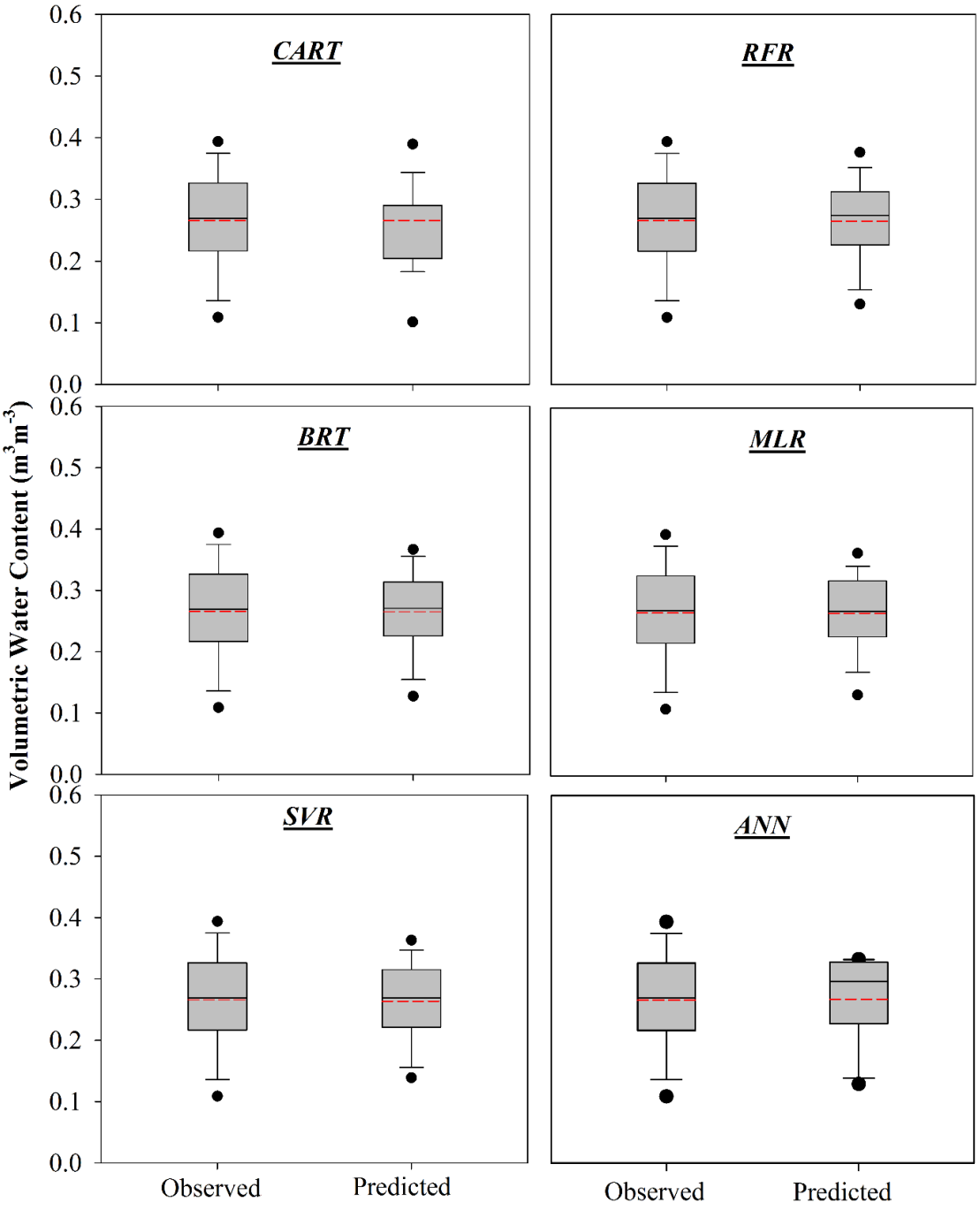

Figure 2.3. Box plots depicting the spread of observed and predicted soil moisture (m^3 m^{-3}) during the testing phase for six different machine learning models. The box shows the interquartile range (25^{th}-75^{th} percentile). The whiskers extend from 5^{th} to 95^{th} percentile values. The solid line inside the box shows the median value (50^{th} percentile) and the dashed line represents the mean value of the observed soil moisture during testing phase. Algorithms included classification and regression trees (CART), multiple linear regression (MLR), random forest regression (RFR), support vector regression (SVR), boosted regression trees (BRT), and artificial neural networks (ANN).

The SVR model provides an appropriate choice of kernels that allow non-separable data in original space to become separable in the feature space. This subsequently helps to obtain non-linear algorithms from algorithms previously restricted to handling linearly separable data sets (Karandish and Simunek, 2016; Bray and Han, 2004). Similar results of SVR model performance over other machine learning models was also observed by Karandish and Simunek (2016) in soil water content prediction under water stress condition; Pal and Mather (2003) for land cover classification; and Gill et al., (2006) for soil moisture prediction in Southwestern Oklahoma. The CART model performed better than the MLR with the formation of binary tree using binary recursive partitioning that yields the maximum reduction in the variability of the response variable (Stewart, 1996). The soil moisture estimation using CART model has also been successfully reported by Han et al. (2018).

Importance of predictor variables

The predictors' variables importance was determined for the three tree-based machine learning models. The variable importance for CART, RFR and BRT models are separated based on the weather station variables (5) and crop field variables (8) (Figure 2.4). The CART and BRT models measure variable importance (as a percentage) by the relative contribution of each variable to the output (crop field soil moisture). However, the RFR measures variable importance based on the percent increase in root mean square error (RMSE) after removing a particular variable and compared with the previous value. The higher the value of percent increase in RMSE, the more important is the variable for soil moisture prediction.

All three models showed soil moisture at the weather station to have the high relative influence for the moisture prediction at nearby fields. This was followed by the four-day cumulative rainfall as PET as the next most important variable. For the CART model, the

weather station VWC, four-day cumulative rainfall and four-day PET had 29%, 20% and 21% relative importance, respectively, in predicting field soil moisture, followed by the weather station's bulk density (8%), the crop field's Ksat (6%) and the weather station's Ksat (4%). Similarly, the BRT model showed that the weather station VWC, four-day cumulative rainfall and four-day PET had the highest influence at 32%, 32%, and 16%, respectively. This was followed by the crop field's clay content and Ksat, which both contributed 4%. Among the CART and BRT models, the weather station VWC, rainfall and PET were the dominant variables for predicting nearby crop field soil moisture. For the RFR model, the four-day cumulative rainfall, four-day PET, weather station VWC and crop field's Ksat had 43%, 40%, 39% and 38% increase in the RMSE, respectively, when they were left out of the model, indicating their relative importance. The soil sand, silt and clay contents had 29%, 30% and 34% increase in RMSE, respectively, while the remaining variables ranged from 23% to 28%.

Another way to evaluate variable importance is by the loss of MAE, which is presented in Figure 2.5. Similar to the previous measures of variable importance, all models showed weather station VWC, four-day cumulative rainfall and PET as the variables with highest importance followed by the silt and clay content from the crop field. The range of loss in MAE was highest in the RFR model (1.6-2.5) as compared to the other models (1.2-1.6).

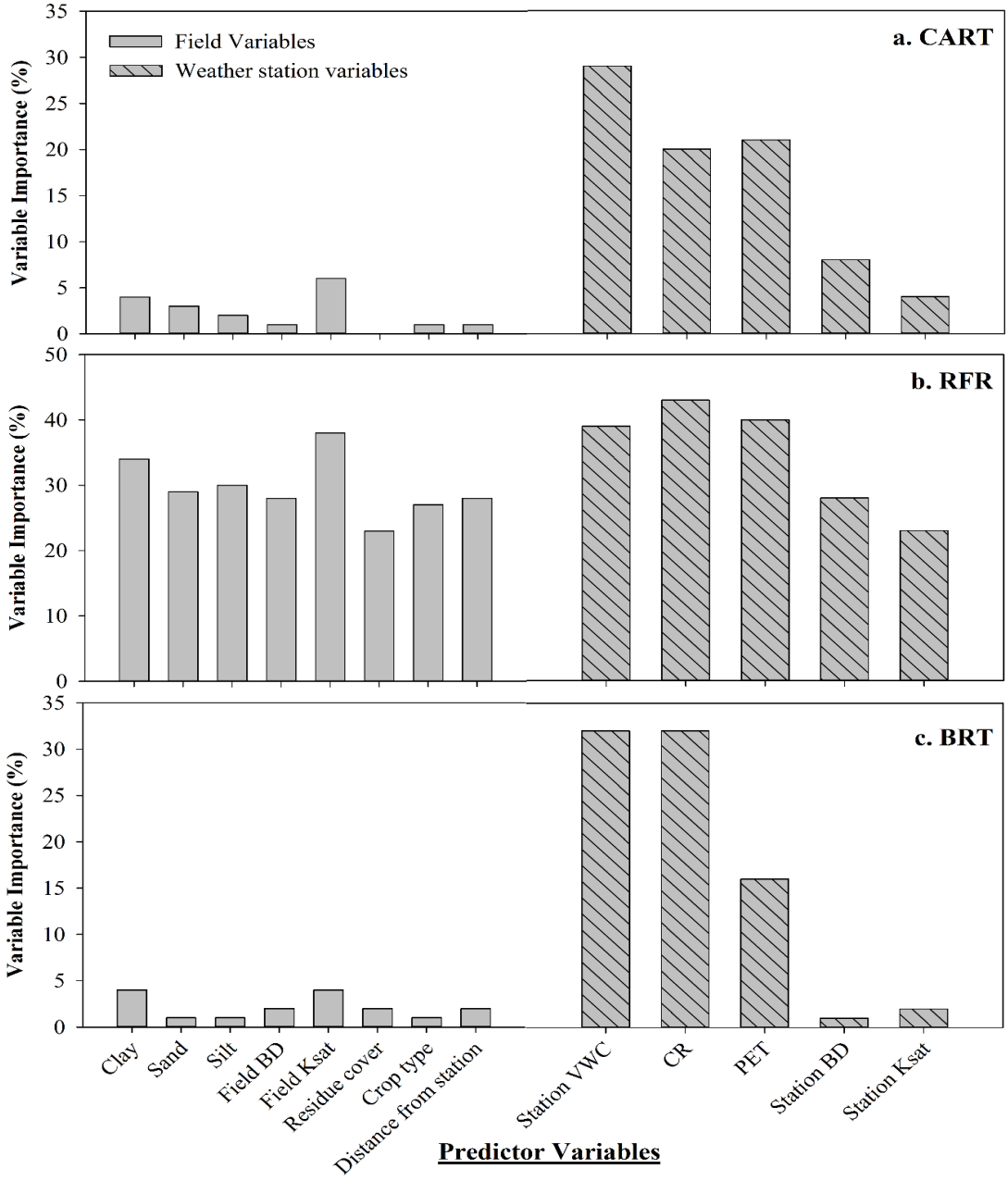

Figure 2.4. Variable importance for three tree-based model types: a. classification and regression trees (CART), calculated as the relative influence (%); b. random forest regression (RFR), calculated as the increase in mean squared error (MSE) (%) and c. boosted regression trees (BRT), calculated as the relative influence (%). Out of 13 predictor variables first 8 represents field and remaining 5 represents station variables. As the calculation of variable importance differs among CART, RFR and BRT, only the ranking of the variables can be compared, but not the absolute values.

These findings are in accordance with Araya et al. (2020), where precipitation was one of the top four important variables for moisture prediction in grassland catchment area of

California, and with Revermann et al. (2016) study where precipitation related variables had high influence on the soil moisture. This is expected since rainfall, and evapotranspiration, have a direct influence on the soil moisture (Brocca et al., 2007; Cosh et al., 2004; Ziadat and Taimeh, 2013). Our study also reveals that Ksat of both the nearby crop field and at the weather stations were the next most influential variables for predicting soil moisture. This is due to Ksat being a governing property for water flows in the soil (Zhang, 1997) and well known to have high spatial variability (Upchyrch et al. 1988). Additionally, soil particle sizes in the crop fields are an important variable since these govern pore sizes and their ability to retain water in the field (Li et al., 2014; Manns et al., 2014). Other variables under this study (i.e., residue cover, crop type and distance from station) showed little influence on the ability to accurately predict soil moisture by these machine learning models, which corroborates Araya et al., (2020), Kravchenko et al. (2011), and McIsaac et al. (2010) studies.

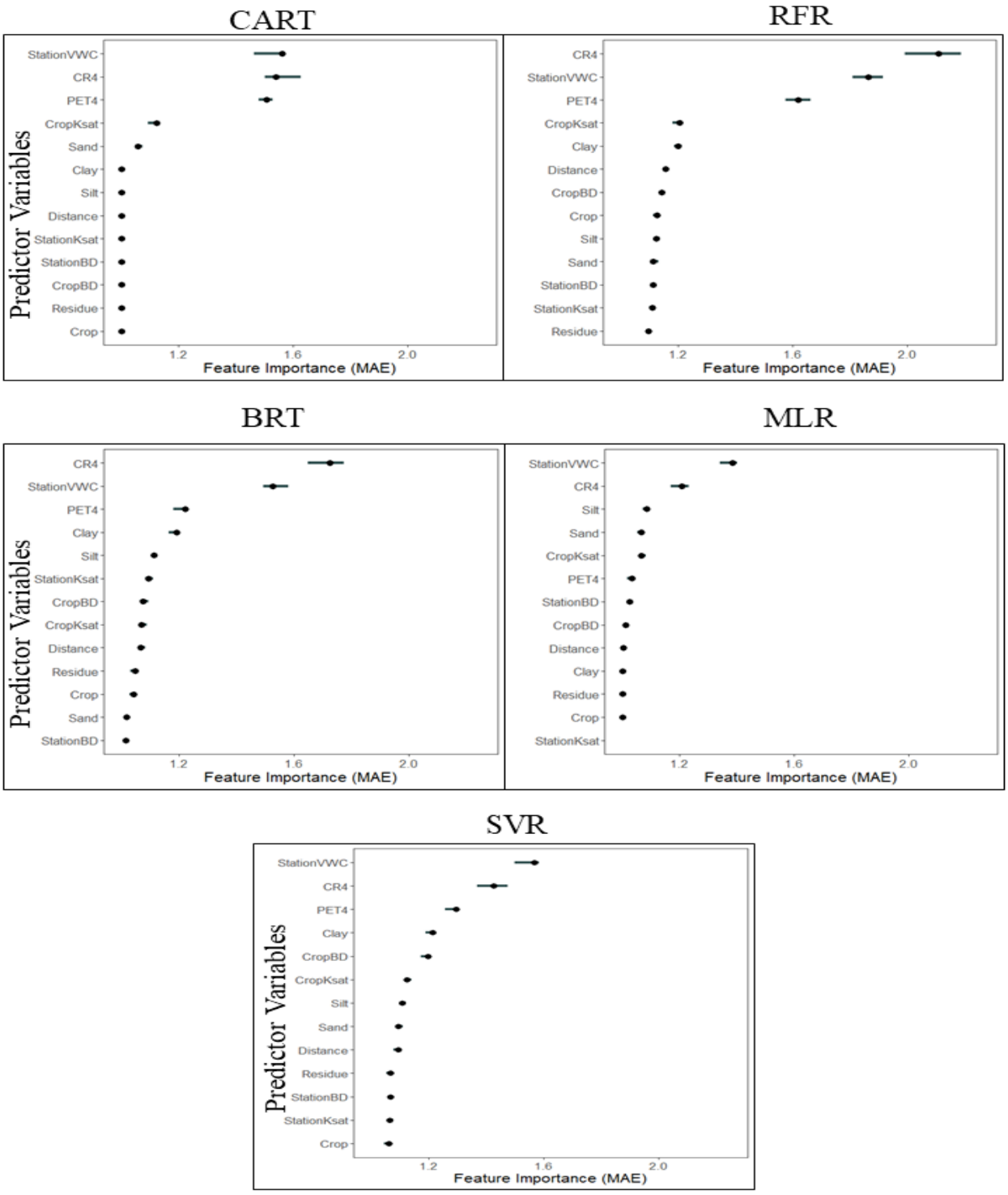

Figure 2.5. Feature importance of predictor variables for five different machine learning model based on the loss of mean absolute error (MAE) along with 5th to 95th percentile values. Algorithms included classification and regression trees (CART), multiple linear regression (MLR), random forest regression (RFR), support vector regression (SVR), and boosted regression trees (BRT).

Accumulated local effect (ALE) of predictor variables

The ALE of predictor variables on predicting soil moisture with the various machine learning algorithms were evaluated graphically (Figure 2.6). The variables previously determined to have a high relative influence (i.e., importance) were graphed, which included the weather station's VWC, four-day cumulative rainfall, four-day cumulative PET and Ksat. The ALE plots show the predicted soil moisture generally increase with higher values of weather station VWC. The CART model showed stationary ALE values except for a sharp increase at 0.25 m^3 m^{-3} in the weather station VWC. This sharp increase in ALE is likely due to the single tree structure of the CART model. In contrast, the RFR model (an ensemble of thousands of trees) showed a gradual sigmoidal increase in the ALE values with station VWC values. This same general trend was observed in the BRT and SVR models.

The ALE for predicting the crop field VWC also increased with the four-day cumulative rainfall (Figure 6). All four models showed a similar trend with the steepness of the ALE slope tending to dissipate with higher weather station VWC and a tendency for the ALE to flatten between 10-25 mm of cumulative rainfall. For the four-day cumulative PET, the ALE was generally stationary until 23 mm of PET, after which the ALE decreased as the PET increased. The only exceptions were slight irregularities in the ALE near 25 mm of PET. This showed that cumulative PET less than 23 mm has no effect on the moisture prediction and have inverse effect on the predicted soil moisture. In this study, the ALE plots showed that station Ksat have minimal effect on the predicted soil moisture. Although there was a slight decrease in predicted soil moisture between 1.5-1.7-inch hr^{-1} in RFR and BRT models, the other models showed only stationary AEL values at zero.

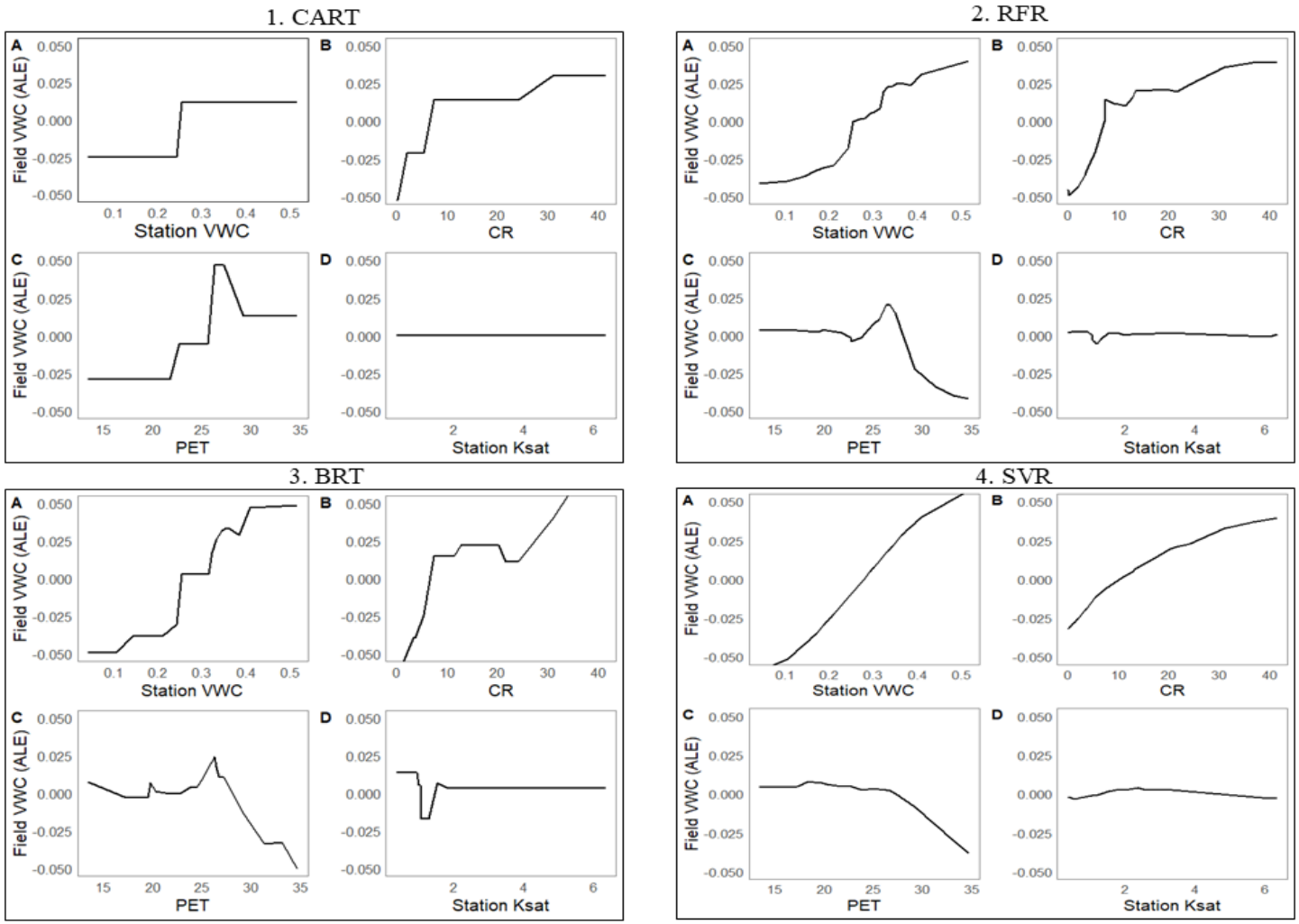

Figure 2.6. Accumulated local effect plots for (a) weather station VWC, (b) four-day cumulative rainfall, (c) four-day cumulative potential evapotranspiration (PET) and (d) weather station saturated hydraulic conductivity (Ksat) under four machine algorithms 1. Classification and regression trees (CART), 2. Random forest regression (RFR), 3. Boosted regression trees (BRT) and 4. Support vector regression (SVR) for model training datasets.

Overall, the cumulative rainfall, PET and weather station VWC are the most important predictor variables for soil moisture prediction in nearby crop fields, which also contain dynamic local effects. For instance, there were particular points for each of predictor variables (i.e., 0.25 m^3 m^{-3} for the weather station VWC, 10 mm for the cumulative rainfall, and 23 mm for PET) which led to large changes in the crop field soil moisture predictions. In Araya et al. (2020), similar dynamics were observed among predictor variables (topography, curvature and flow accumulation) on soil moisture estimates using ALE plots for a BRT model. Similarly, other studies have showed the significant effect of cumulative rainfall and PET on predicting soil moisture, for example, Entekhabi and Rodriguez-Iturbe, (1994), Pan et al., (2003), and Brocca et al., (2013).

Conclusion

Machine learning algorithms can be used effectively in predicting field soil moisture. These algorithms are based on different principles (regression trees, kernels, and regression) and results in different levels of effectiveness in prediction. Successful soil moisture prediction involves establishing the effect of each variable on the output (variable importance). The RFR, BRT and SVR predictions performed better than the remaining algorithms based on high correlations, low RMSE and MAE, during model validation using an independently derived dataset. The weather station variables (station soil moisture, four-day cumulative rainfall, and PET) were relatively more influential than the soil and crop variables for predicting field soil moisture in the nearby plots.

In summary, the following conclusions can be drawn from this study:

- RFR, BET and SVR outperformed other models in soil moisture prediction based on the r^2, RMSE and MAE values.

- RFR showed the highest r^2 (0.72), and lowest MAE (0.034 m^3 m^{-3}) and RMSE (0.045 m^3 m^{-3}).
- RFR, CART and BRT showed the weather station soil moisture, four-day cumulative rainfall and PET have a high influence compared to soil and crop factors on predicting soil moisture in nearby crop fields.
- ALE plots showed the weather station moisture, four-day cumulative rainfall and PET as the most important predictor variables for soil moisture prediction close to crop fields.

References

Achieng, K.O., 2019. Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models. Comput. Geosci. 133, 104320.

Ahmad, S., Simonovic, S.P., 2005. An artificial neural network model for generating hydrograph from hydro-meteorological parameters. J. Hydrol. 315(1-4), 236-251.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7(12), 16398–16421, doi:10.3390/rs71215841.

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., Dehnavi, S., 2017. Temperature-Vegetation-soil Moisture Dryness Index (TVMDI). Remote Sens. Environ. 197, 1-14. http://dx.doi.org/10.1016/j.rse.2017.05.026

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. B. 82(4), 1059-1086.

Araya, S.N., Fryjoff-Hung, A., Anderson, A., Viers, J.H. and Ghezzehei, T.A., 2020. Advances in Soil Moisture Retrieval from Multispectral Remote Sensing Using Unmanned Aircraft Systems and Machine Learning Techniques. Hydrol. Earth Syst. Sc. 1-33.

Asefa T., Kemblowski, M., McKee, M., Khalil, A., 2006. Multi-time scale stream flow predictions: the support vector machines approach. J. Hydrol. 318, 7–16.

Bray, M., Han, D., 2004. Identification of support vector machines for runoff modeling. J. Hydrol. 6 (4), 265–280.

Breiman, L., 2001. Random Forest, Mach. Learn. 45(1), 5–32, doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. CART. Classification and Regression Trees, Monterey, CA: Wadsworth and Brooks/Cole.

Brocca L, Morbidelli, R., Melone, F., Moramarco, T., 2007. Soil moisture spatial variability in experimental areas of Central Italy. J. Hydrol. 333, 356–373. https://doi.org/10.1016/j.jhydrol.2006.09.004

Brocca, L., Moramarco, T., Melone, F., Wagner, W., 2013. A new method for rainfall estimation through soil moisture observations. Geophys. Res. Lett. 40(5), 853-858.

Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., Xue, X., 2019. Research on soil moisture prediction model based on deep learning. Plos one 14(4), 0214508.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning. 161-168.

Clapcott, J., Goodwin, E., Snelder, T., 2013. Predictive models of benthic macro-invertebrate metrics. Cawthron Report No. 2301, 35

Cortes, C., Vapnik, V.N., 1995. Support-Vector Networks, Mach. Learn. 20(3), 273-297 doi:10.1023/A:1022627411411, 1995.

Cosh, M.H., Stedinger, J.R., Brutsaert, W., 2004. Variability of surface soil moisture at the watershed scale. Water Resour. Res. 40, W12513.

Daigh, A.L.M., DeJong-Hughes, J., Gatchell, D.H., Derby N.E., Alghamdi R., Leitner Z.R., Wick, A., Acharya, U., 2019. Crop and soil responses to on-farm conservation tillage practices in the Upper Midwest. Agric. Environ. Lett. 4(1), 1-5 doi:10.2134/ael2019.03.0012

De'Ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology, 88(1), 243-251.

Dibike Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and application. J. Comput. Civil Eng. 15(3), 208-16.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Advances in neural information processing systems, 9, 155-161.

Elith, J., Leathwick, J. R. and Hastie, T. 2008. A working guide to boosted regression trees, J. Anim. Ecol., 77(4), 802–813, doi:10.1111/j.1365-2656.2008.01390.x.

Entekhabi, D., Rodriguez-Iturbe I., 1994. Analytical framework for the characterization of the space-time variability of soil moisture. Adv. Water Resour. 17, 35-45.

Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. 29(5), 1189-1232, 605 doi:10.1017/CBO9781107415324.004.

Friedman, J.H., 2002. Stochastic gradient boosting, Comput. Stat. Data Anal., 38(4), 367-378. doi:10.1016/S0167- 9473(01)00065-2.

Gee, G.W., Bauder, J.W., 1986. Particle Size Analysis. In: Methods of Soil Analysis, Part A. Klute (ed.). 2 Ed., Vol. 9. Am. Soc. Agron., Madison, WI, 383-411.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. As. 42(4), 1033-1046.

Gorthi, S., Dou, H., 2011. Prediction models for the estimation of soil moisture content. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference 54808, 945-953.

Greenwell, B.M., 2017. An R Package for Constructing Partial Dependence Plots. R J., 9(1), 421.

Grimes, D.I.F., Coppola, E., Verdecchia, M., Visconti, G., 2003. A neural network approach to real-time rainfall estimation for Africa using satellite data. J. Hydrometeorol. 4(6), 1119-1133.

Gumiere, S.J., Camporese, M., Botto, A., Lafond, J.A., Paniconi, C., Gallichand, J., Rousseau, A.N., 2020. Machine Learning vs. Physics-Based Modeling for Real-Time Irrigation Management. Fro. Water, 2, 8. https://doi.org/10.3389/frwa.2020.00008

Günther, F., Fritsch, S., 2010. Neuralnet: Training of neural networks. The R journal, 2(1), 30-38.

Hamman, B., Egil, D.B., Koning, G., 2002. Seed vigor, soilborne pathogens, pre-emergent growth, and soybean seeding emergence. Crop Sci. 42, 451-457. https://doi.org/10.2135/cropsci2002.0451

Han, J., Mao, K., Xu, T., Guo, J., Zuo, Z., Gao, C., 2018. A soil moisture estimation framework based on the CART algorithm and its application in China. J. Hydrol. 563, 65-75.

Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., McKee, M., Assessment of surface soil moisture using high-resolution multi-spectral imagery and artificial neural networks, Remote Sens. 7(3), 2627–2646, doi:10.3390/rs70302627, 2015.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., 2017. SoilGrids250m: Global gridded soil information based on machine learning. Plos one, 12(2), e0169748.

Huang, C., Li, L., Ren, S., Zhou, Z., 2010, Research of soil moisture content forecast model based on genetic algorithm BP neural network. In International Conference on Computer and Computing Technologies in Agriculture. Springer, Berlin, Heidelberg. 309-316

Igwe, C.A., 2005. Soil physical properties under different management systems and organic matter effects on soil moisture along soil catena in southeastern Nigeria. Trop. Subtrop. Agroecosyst. 5(2), 57-66.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer New York, New York, NY.

Kalra, A., Ahmad S., 2009. Using oceanic–atmospheric oscillations for long lead time streamflow forecasting. Water Resour. Res. 45, W03413. doi:10,1029/ 2008WR006855.

Karandish, F., Simunek, J., 2016. A comparison of numerical and machine-learning modeling of soil water content with limited input data. J. Hydrol. 543, 892-909.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab: An S4 Package for Kernel Methods in R (version 0.9-25), J. Stat. Softw. 11(9), 1–20 [online] Available from: http://www.jstatsoft.org/v11/i09/,

Kashif Gill, M., Kemblowski, M.W., McKee, M., 2007. Soil moisture data assimilation using support vector machines and ensemble Kalman filter. J. Am. Water Resour. As. 43(4), 1004-1015.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma, 339, 40-58.

Kravchenko, A.N., Wang, A.N.W., Smucker, A.J.M., Rivers, M.L., 2011. Long-term differences in tillage and land use affect intra-aggregate pore heterogeneity. Soil Sci. Soc. Am. J. 75, 1658–1666. doi:10.2136/sssaj2011.0096

Laguardia, G., Niemeyer, S., 2008. On the comparison between the LISFLOOD modelled and the ERS/SCAT derived soil moisture estimates. Hydrol. Earth Syst. Sci., 12(6), 1339-1351.

Li, B., Yang, G., Wan, R., Dai, X., Zhang, Y., 2016. Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. Hydrol. Research, 47(S1), 69-83.

Li, T., Hao, X.M., Kang, S.Z., 2014. Spatiotemporal variability of soil moisture as affected by soil properties during irrigation cycles. Soil Sci. Soc. Am. J. 78(2), 598–608. https://doi.org/10.2136/sssaj2013.07.0269

Liaw, A., Wiener, M., 2002. Classification and regression by RandomForest. R news, 2(3), 18-22.

Liaw, A., Wiener, M., 2015 RandomForest: Breiman and Cutler's random forests for classification and regression. R package version 4, 6-10, 2015

Liong, S.Y., Sivapragasam, C., 2002. Flood stage forecasting with support vector machines 1. J. Am. Water Resour. As. 38(1), 173-186.

Manns, H. R., Berg, A. A., Bullock, P. R., McNairn, H. 2014. Impact of soil surface characteristics on soil water content variability in agricultural fields. Hydrological Processes, 28(14), 4340–4351. https://doi.org/10.1002/hyp.10216

Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A data mining system for real time soil moisture prediction. Procedia Engineer. 181, 837-844.

McIsaac, G.F., David, M.B., Mitchell, C.A., 2010. Miscanthus and switchgrass production in central Illinois: Impacts on hydrology and inorganic nitrogen leaching. J. Environ. Qual. 39, 1790–1799. doi:10.2134/jeq2009.0497

Molnar, C., 201. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. [online] Available from: https://christophm.github.io/interpretable-ml-book/

Naghibi, S.A., Pourghasemi, H.R., Dixon, B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188(1), 44.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial, Front. Neurorobot., 7, 1–11, doi:10.3389/fnbot.2013.00021.

NOAA/NCEI, 2020. National Oceanic and Atmospheric Administration/National Centers for Environmental Information. https://www.ncdc.noaa.gov/ (Data retrieved on 7/15/2020)

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4(1), 1-22.

Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. Q. Rev. Biol. 83(2), 171-193.

Pal, M., Mather, P.M., 2003. Support Vector Classifiers for Land Cover Classification. Map India 2003, Image processing and interpretation. Available at http://www.gisdevelopment.net/technology/rs/pdf/23.pdf. Accessed in 2004.

Pan, F., Peters-Lidard, C.D., Sale, M.J., 2003. An analytical method for predicting surface soil moisture from rainfall observations. Water Resour. Res. 39(11), 1-12.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A. Ma. P. Scis., 193(1032), 120-145.

Priddy, K.L., Keller, P.E., 2005. Artificial neural networks: an introduction Vol. 68. SPIE press.

Qiao, X., Yang F., Xu, X., 2014. The prediction method of soil moisture content based on multiple regression and RBF neural network. Ground Penetrating Radar (GPR) 2014 15th International Conference, 140-143.

R Development Core Team. 2020. R: A Language and Environment for Statistical Computing, [Online] Available from: https://www.r-project.org/, 2020.

Revermann, R., Finckh, M., Stellmes, M., Strohbach, B.J., Frantz, D., Oldeland, J., 2016. Linking land surface phenology and vegetation-plot databases to model terrestrial plant α-diversity of the Okavango Basin. Remote Sens. 8(5), p.370.

Reynolds, S.G., 1970. The gravimetric method of soil moisture determination Part IA study of equipment, and methodological problems. J. Hydrol., 11(3), 258-273.

Samadi, M., Jabbari, E., Azamathulla, H.M., 2014. Assessment of M5 model tree and classification and regression trees for prediction of scour depth below free overfall spillways. Neural Comput. Appl. 24(2), 357-366.

Sanuade, O.A., Hassan, A.M., Akanji, A.O., Olaojo, A.A., Oladunjoye, M.A., Abdulraheem, A., 2020. New empirical equation to estimate the soil moisture content based on thermal properties using machine learning techniques. Arab. J. Geosci. 13, 1-14.

Schaap, M.G., Leij, F.J., vanGenuchten, M. Th., 2001. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol., 251, 163-176.

Simunek, J., Sejna, M., Saito, H., Sakai, M., vanGenuchten, M.Th., 2008. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media version 4.0. Department of Environmental Sciences, University of California Riverside, California.

Stewart, J.R., 1996. Applications of Classification and Regression Tree Methods in Roadway Safety Studies. In Transportation Research Record 1542, TRB, National Research Council, Washington, D.C. 1-5.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23(6), pp.2615-2635.

Twarakavi, N.K., Misra, D., Bandopadhyay, S. 2006. Prediction of arsenic in bedrock derived stream sediments at a gold mine site under conditions of sparse data. Nat. Resour. Res. 15(1), 15-26.

Upchurch, D.R., Wilding, L.P., Hatfield, J.L., 1988. Methods to evaluate spatial variability. In Reclamation of surface-mined land. 201-229

USDA. 2020. United States Department of Agriculture, International Production Assessment Division. Metadata for crops at different growth stage.

https://ipad.fas.usda.gov/cropexplorer/description.aspx?legendid=312 (data retrieved on 5/8/2020)

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1.

Venables W.N., Ripley, B.D., 2002. Modern applied statistics with S. Statistics and computing. New York: Springer.

Yang, H., Huang, K., King, I., Lyu, M.R., 2009. Localized support vector regression for time series prediction. Neurocomputing, 72(10-12), 2659-2669.

Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O., Lee, K.K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 396(1-2), 128-138.

Zaman, B., Mckee, M., 2014. Spatio-temporal prediction of root zone soil moisture using multivariate relevance vector machines, Open J. Mod. Hydrol. 4(3), 80–90, doi:dx.doi.org/10.4236/ojmh.2014.43007, 2014.

Zaman, B., McKee, M., Neale, C.M.U., 2012. Fusion of remotely sensed data for soil moisture estimation using relevance vector and support vector machines, Int. J. Remote Sens., 33(20), 6516-6552, doi:10.1080/01431161.2012.690540.

Zhang, H.X., Yang, J., Fang, X.Y., Fang, J., Feng, C., 2008. Application of time series analysis in soil moisture forecast. Res. Soil Water Conser. 15(4), 82-84.

Zhang, R., 1997. Determination of soil sorptivity and hydraulic conductivity from the disk infiltrometer. Soil Science Society of America Journal, 61(4), 1024-1030.

Zhang, W., Yuan, S., Hu, N., Lou, Y., Wang, S., 2015. Predicting soil fauna effect on plant litter decomposition by using boosted regression trees. Soil Biology and Biochemistry, 82, 81-86.

Zhou, L., 2007. Study on estimation of soil-water content by using Soil-Water Dynamics Model [J]. Water Saving Irrigation, 3, 10-13.

Ziadat, F.M., Taimeh, A.Y., 2013. Effect of rainfall intensity, slope, land use and antecedent soil moisture on soil erosion in an arid Environment. Land Degrad. Dev. 24(6), 582-590.

CHAPTER III. AN INTEGRATED RANDOM FOREST–OPTRAM ALGORITHM PERFORMED BETTER THAN VEGETATIVE INDICES AND OPTRAM FOR MAPPING SURFACE SOIL MOISTURE FROM LANDSAT 8 IMAGES

Abstract

Remote sensing tools have been extensively used for large scale soil moisture mapping in recent years using Landsat satellite images. Rainfall, soil clay percent, standardized precipitation index play key roles in the moisture content of the crop field. Large scale soil moisture prediction required to process large amounts of data and machine learning algorithms effectively used. The objective of this study was to (i) obtain representative soil moisture dataset across agricultural geographic region with complex mosaic of crop species and soil management, (ii) calculate and determine the effectiveness of vegetation indices in predicting surface soil moisture, (iii) predict surface soil moisture from satellite images using (Optical Trapezoid Model (OPTRAM), and (iv) evaluate if the OPTRAM predictions can be improved by incorporating weather station, soil and crop data with a random forest machine learning algorithm. ENVI® platform was used to create vegetation indices maps and google earth engine (GEE) was used to prepare OPTRAM maps. Random forest regression was performed on the R-software platform. Results showed a very weak relationship between the vegetation indices and surface soil moisture content where r^2 and slopes were < 0.10 and < 0.20 respectively. OPTRAM soil moisture when compared with in situ surface moisture showed weak relationship with regression values < 0.2. Surface soil moisture was then predicted using random forest regression using OPTRAM moisture values, rainfall, standardized precipitation index (SPI) and percent clay showed high goodness of fit (r^2=0.69) and low root mean square error (RMSE = 0.053 m^3 m^{-3}). This showed that rainfall, SPI and percent clay are important infield characteristics that can be

used along with OPTRAM moisture values to improve soil moisture prediction in the agricultural field. The surface soil moisture prediction model was developed using OPTRAM values, rainfall, SPI and percent clay map for the Red River Valley of North.

Key words: OPTRAM, random forest, vegetation indices, soil moisture prediction

Introduction

Remote sensing has been used as an advance tool for agricultural interpretation (Lillesand et al., 2008). A prime area of research in agriculture is the in-field variability of plant water stress across large scales, which directly relates to in-field soil moisture (Bastiaanssen et al., 2000). Estimation of surface soil moisture provides farmers with key information on water stress, which aids crop yield projective, assessment of drought and excessive water conditions, and informs water management practices (e.g., irrigation, drainage) (Penuelas et al., 1993; Tucker, 1980). Remote sensing can be effectively used to estimate soil moisture because soil optical reflection and thermal emissions are highly correlated with soil moisture (Zeng et al., 2016; Zhang and Zhou, 2016). Remotely sensed visible and thermal infrared wavelengths provide more information when combined, rather than alone (Zhang et al., 2014). However, spatial and temporal information are needed to produce precise and accurate estimates (Zhang and Zhou, 2016). Remote sensing techniques provide tools for mapping soil moisture at large spatial and temporal scales (Das and Paul, 2015). Several mathematical models using remotely sensed data have been developed to estimate soil moisture using satellite optical image datasets, such as Landsat and sentinel that are freely available. Remote sensing provides time and cost-effective data, which are valuable for soil moisture estimation at regional and national level using various techniques (Verstraeten et al., 2008; Wang and Qu, 2009; Peng et al., 2017). There are numerous methods used in remote sensing to estimate soil moisture. All these methods are based on the

type of image used, such as microwaves (Francois, 2000), thermal, visible and infrared (Sobrino and Raissouni, 2000), radiometric behavior of infrared waves (Levit et al. 1990), and the use of relationships between surface temperature and fractional vegetation cover to estimate soil moisture (Dupigny-Giroux and Lewis, 1999). Vegetation indices can be misleading for estimating soil moisture due to the time lag between changes in soil moisture and corresponding changes in the vegetation indices (Sandholt et al. 2002; Tadesse et al., 2005)

Vegetation indices

One common approach to estimate soil moisture across landscapes has been to derive moisture from proxy measurements like vegetation indices derived from satellite imagery. For instance, investigators have used reflectance data to estimate soil moisture by the greenness and water content in the leaves of crop canopies, such as Normalized Difference Vegetation index (NDVI) (Rouse et al., 1973) and Normalized Difference Water Index (NDWI) (Gao, 1996). NDVI is a routinely produced and used product for indicating vegetation water content, soil moisture, and crop yield prediction. NDVI has limitations for estimating soil moisture because each crop species has its own unique relation with chlorophyll content, and a decrease in chlorophyll does not imply low soil moisture (Ceccato et al., 2001). Jackson et al. (2004) found NDWI as a superior index to NDVI for estimating soil moisture. For examples, in maize, they reported that NDVI and NDWI have a root mean square error (RMSE) of 0.735 and 0.576 kg m^{-2} with a bias of 0.336 and -0.010 kg m^{-2}, respectively. Similarly, in soybean, RMSE of 0.203 and 0.171 kg m^{-2} with bias of 0.071 and -0.015 kg m^{-2} were observed for NDVI and NDWI, respectively. NDWI determines the water volume per leaf, which mostly depends on the leaf area index and crop canopy and not necessarily soil moisture content. Other vegetation indices such as Normalized Difference Moisture Index (NDMI), Enhanced Vegetation Index (EVI),

Atmospherically Resistant Vegetation Index (ARVI), and Structure Insensitive Pigment Index (SIPI) have been used to established relationships with soil moisture using different bands (green, blue, red, infrared, short-wave infrared). These indices have been successful in some region-specific cases, but not generally across all cases and lacks wide application (Jackson et al., 2004).

Physically-based models

Physically-based models are commonly preferred over the empirical vegetation indices, but they also have their own limitations due to various confounding factors and the need for specific parameters. For instance, land surface temperature (LST) has been used as key parameter that relates with surface energy and water balance at local and large scales and helps in monitoring climate, vegetation and hydrological cycles (Liang, 2004). Energy is exchanged between land surface and atmosphere when soil moisture evaporates into atmosphere. This demonstrates that soil moisture not only depends on rainfall but also on soil surface temperature. The triangle (NDVI-LST) method is a widely used model (Rahimzadeh-Bajgiran et al., 2013; Shafian and Mass, 2015; Sun, 2016) for estimating soil moisture, but it has two major limitations. The first limitation is the requirement of thermal data, which is not applicable to all types of satellite images. The second limitation is that LST is affected by wind speed, air temperature, and humidity.

To overcome these two limitations, Sadeghi et al. (2017) proposed the physically-based Optical Trapezoidal Model (OPTRAM) for soil moisture estimation. This model uses Short Wave Infrared (SWIR) transformed reflectance (STR) instead of LST. This results in the STR-NDVI space to remain nearly time invariant because reflectance is a function of only the surface properties and not the ambient atmospheric condition as LST. The OPTRAM model forms a

trapezoid space using both NDVI as a measure of vegetative fraction and the STR to establish the linear relationship for dry and wet edges that helps to estimate soil moisture. STR does not significantly change with ambient atmospheric condition and can be universally parameterized for a given location. The combination of vegetation indices and STR can provide useful information for detection of spatial and temporal distribution of soil moisture. Additionally, researchers have used historical meteorological records to develop time-invariable coefficients, which are used with remote sensing data such as LST and STR to estimate soil moisture (Leng et al., 2016). This is important in improving the spatial and temporal resolution of surface soil moisture information for precision agriculture. There is great potential of using optical/thermal satellite images and meteorological data to develop an all-weather soil moisture model (Leng et al., 2017).

However, soil moisture mapping using satellite image processing may not provide accurate values in all cases, because field soil moisture is dependent on a suite of complex factors (canopy, crop type and growth stage, crop residue, soil type). In addition to field conditions, rainfall is another important factor affecting soil moisture content. There has been significant work in establishing relation between rainfall and soil moisture, so that rainfall data can be used to predict soil moisture. Brocca et al. (2013) found four-day cumulative rainfall can be effectively used to predict soil moisture with correlation coefficient values of 0.8 and root mean square error (RMSE) of 1.36 mm day^{-1}. Standardized Precipitation Index (SPI) is used to measure precipitation change over time at different time scale level (1, 3, 6, 12 months) and is a valuable tool for indicating meteorological drought using past rainfall patterns.

Potential for machine learning to overcome challenges with OPTRAM

Despite the progress for estimating high-resolution soil moisture across large areas, an integrated approach to predict soil moisture with the use of meteorological data, crop and field data and remote sensing images has been lacking (Yu et al., 2017). Additionally, some studies have reported poor performance of the OPTRAM model for specific regions. For instance, Babaeian et al. (2018) observed r2 values from 0.01 to 0.49 and RMSE ranges from 0.05 to 0.08 m3 m-3 soil moisture in the United States. Yadav et al. (2019) reported accuracies of less than 10% when compared with other models (thermal) for the Lalitpur district of India. Yadav et al. (2019) concluded this reduced accuracy might be due to lack of penetration of optical bands in the fully or densely covered vegetation. Similarly, Chen et al. (2020) evaluated OPTRAM soil moisture estimates using MODIS data and observed r^2 values of 0.10 to 0.50 with RMSE's from 0.05 to 0.13 m^3 m^{-3}, respectively. They suggested the poor performance between the OPTRAM soil moisture estimates and in situ soil moisture might be due to rough image resolution and heterogenous terrains (Chen et al., 2020).

These studies characterized the spatial distribution of surface soil moisture solely by using remote sensing. Very few studies have been conducted to incorporate climatic conditions and geographical or field and crop characteristics (Vicente-Serrano et al., 2004). As part of the surface-atmosphere energy flux, parameters such as precipitation, LST and ET can substantially modify soil moisture across the landscape (Alfieri et al., 2008; Hu et al., 2017). Recently, machine learning techniques have become useful in predicting soil moisture under different field conditions. Machine learning techniques have the advantage and ability to learn and approximate complex non-linear mapping while requiring no assumptions on the distribution of the data. Machine learning techniques can also integrate data from different sources with poorly-defined

or unknown probability density functions (Ali et al., 2015). For example, Ali et al. (2015) and Paloscia et al. (2008) showed machine learning techniques (e.g., Artificial Neural Network, Support Vector Regression) can outperform other parametric approaches for estimating soil moisture and improved their performance with an increasing number of observed datasets. Therefore, integrating meteorological data from weather Mesonet and field characteristics (soil and crop data) along with OPTRAM soil moisture values in a machine learning algorithm has the potential to be a valuable tool in mapping high-resolution soil moisture across large areas.

Study objectives

The objectives of this study were to (1) obtain a representative surface soil moisture dataset across an agricultural geographic region with a complex mosaic of crop species and soil management on dates aligned with the Landsat 8 satellite, (2) calculate and determine the effectiveness of vegetation indices in predicting surface soil moisture, (3) predict surface soil moisture from the satellite images using OPTRAM, and (4) evaluate if the OPTRAM predictions can be improved by incorporating weather station, soil, and crop data with a Random Forest machine learning algorithm.

Material and Methods

Study area

The study area covered eight counties of North Dakota and seven counties of Minnesota that lies on either side of Red River of North (Latitude = 45.96 to 48.99 and Longitude = - 95.51 to - 98.01). Surface soil moisture was collected for 2019 crop growing season. Maize (*Zea mays* L.), soybean (*Glycine max* (L.) Merr.), wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), sugar beet (*Beta vulgaris*), canola (*Brassica napus*), sunflower (*Helianthus annuus* L.), potato (*Solanum tuberosum* L.), dry beans (*Phaseolus vulgaris*), and oats (*Avena sativa* L.) are

major crops grown in the Red River Valley of North (RRVN). Twenty-five weather stations (NDAWN, 2020) distributed over counties of MN (n=10) and ND (n=15) were selected for this study that covers the RRVN (Figure 3.1).

Figure 3.1. Location of NDAWN stations in counties of ND and MN around the Red River Valley of North.

Field data collection

Soil samples were collected from the 25 weather stations and three adjacent crop fields within 2 Km of each weather station (i.e., 100 sample collection sites). Samples from all 100 collection sites were obtained at 16 days intervals, coinciding with the Landsat 8 satellite passing days (June 18 to September 29, 2019). The adjacent crop fields near the weather stations were planted with soybean (n = 24), wheat (n = 18), corn (n = 16), sugar beet (n = 6), dry beans (n = 5), oats (n = 2), barley (n = 1), potato (n = 1), canola (n = 1) and alfalfa (n = 1). At each collection site, three composite soil samples (0 – 6 cm) were collected and averaged to get a representative surface soil moisture. Gravimetric water content was calculated using oven drying method and volumetric water content (VWC) was determined by multiplying the bulk density with the gravimetric water content (Reynolds, 1970). Growth stages of crops were monitored following the standard methods by United States Department of Agriculture (USDA, 2020) for each cropped field and recorded every 16 days.

Satellite image processing

Vegetation indices

Multispectral NASA Landsat 8 (L8) satellite images were acquired from the United States Geological Survey (USGS) Earth explorer (URL: https://earthexplorer.usgs.gov/) for the study area. Landsat-8 houses the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS), which image the land surface at 11 spectral bands in the optical and thermal infrared domains with 30 m to 100 m spatial resolution and 16-day temporal resolution. A total of 28 level-1 L8 images acquired in the 2019 growing season were used in this study (Table 3.1). The Fmask algorithm was used to detect and remove clouds and cloud shadows for all Landsat imagery (Zhu et al., 2015).

Radiometric corrections of the multispectral bands were performed using the Fast Line-of-sight Atmospheric Analysis of Spectral Hypercube (FLAASH) toolbox, included in the Environment for Visualizing Image (ENVI) 5.5 software (https://www.harrisgeospatial.com/). The FLAASH atmospheric correction algorithm was chosen due to its better performance compared to algorithms, such as Quick Atmospheric Correction (QUAC) and Dark Object Subtraction (DOS) (Shi et al., 2016). After removing clouds and obtaining the atmospheric corrected bands, a variety of vegetation indices were calculated for each pixel as described below.

Table 3.1. List of the Landsat 8 image (path/row) acquired over the study area covering weather stations in 2019.

Path/Row	Weather station	Date Sampled (2019)
29/28	Campbell, Mooreton, Wahpeton, Fargo Sabin	6/11, 6/27, 7/13, 7/29, 8/14, 8/30, 9/15
30/26	Forest River, Inkster, Warren, Grafton, St. Thomas, Kennedy, Cavalier, Humboldt	6/18, 7/4, 7/20, 8/5, 8/21, 9/6, 9/22
30/27	Leonard, Sabin, Fargo, Ulen, Prosper, Galesburg, Perely, Hillsboro, Ada, Waukon, Mayville, Finley, Eldred, Grand Forks, Forest River, Inkster, Warren,	6/18, 7/4, 7/20, 8/5, 8/21, 9/6, 9/22
31/26	Grafton, St. Thomas, Kennedy, Cavalier, Humboldt, Forest River, Inkster	6/25, 7/11, 7/27, 8/12, 8/28, 9/13, 9/29

Six vegetation indices were calculate using band math toolbox of ENVI software. These indices included Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Normalized Difference Moisture Index (NDMI), Enhanced Vegetation Index (EVI), Structure Insensitive Pigment Index (SIPI) and Atmospherically Resistant Vegetation Index (ARVI). The vegetation indices were calculated using equations that involve different bands (blue, green, red, NIR, SWIR) and ranges (descriptions shown in Table 3.2).

Those images of the vegetation index were processed in ArcGIS pro software and pixel values for each location of soil sampling were extracted from the image. These normalized values of vegetation indices were then plotted against the measured in-situ surface soil moisture (VWC).

Table 3.2. List of vegetation indices along with formula to calculate, and their range for Landsat 8 satellite image used in this study.

Index	Formula for calculation	Range	Reference
Normalized Difference Vegetation Index (NDVI)	NDVI = (NIR - R) / (NIR + R)	-1 to +1; Where +1 represents dense green leaves and -1 represents a likely water body	Tucker, 1979
Normalized Difference Water Index (NDWI)	NDWI = (Green - NIR) / (Green + NIR)	-1 to +1; Where +1 represents extensive deep-water bodies and -1 represents vegetation cover	McFeeters, 1996; Xu, 2006; Sims and Gamon, 2003
Normalized Difference Moisture Index (NDMI)	NDMI = (NIR - SWIR) / (NIR + SWIR)	-1 to +1; Where +1 represents high canopy cover and no water stress and -1 represents low canopy cover to bare soil	Gao, 1996
Enhanced Vegetation Index (EVI)	EVI = 2.5[(NIR - R) / (NIR + 6R - 7.5Blue + 1)]	-1 to +1; healthy vegetation generally falls between values of 0.20 to 0.80	Liu and Huete, 1995
Structure Insensitive Pigment Index (SIPI)	SIPI = (NIR - Blue) / (NIR - R)	0 to 2; healthy green vegetation is from 0.8 to 1.8.	Penuelas et al., 1995
Atmospherically Resistant Vegetation Index (ARVI)	ARVI = (NIR - 2R + Blue) / (NIR + 2R + Blue)	-1 to +1 healthy vegetation generally falls between values of 0.20 to 0.80	Kaufman and Tanre, 1992

Where, wavelengths range for blue band is 0.450-0.515µm; green band is 0.525-0.600µm; red (R) band is 0.600-0.680µm; near infrared (NIR) band is 0.845-0.885µm; and shortwave infrared (SWIR) band is 1.560-1.660µm at Landsat 8 satellite image.

OPTRAM model

The OPTRAM model developed by Sadeghi et al. (2017) to estimate soil moisture is a physically-based trapezoidal space of pixel distribution within the STR-NDVI space (Figure 3.2). The NDVI is normalized difference vegetation index and STR is SWIR transformed reflectance (Sadeghi et al., 2015). The OPTRAM model only uses optical data, which means that no thermal infrared data area is used for retrieving soil moisture. The normalized soil moisture content (*W*) for each pixel was estimated from the dry edge and wet edge parameters as follows:

$$W = \frac{STR - STR_d}{STR_w - STR_d}, \tag{3.1}$$

where, *STR* is the SWIR transformed reflectance, as follows:

$$STR = \frac{(1-R_{SWIR})^2}{2R_{SWIR}}, \tag{3.2}$$

where, STR_d and STR_w are *STR* at the dry (e.g., $\theta_d \sim 0\%$, where θ_d is soil moisture content) and wet (e.g., $\theta_s \sim 100\%$, where θ_s is soil moisture content) states, respectively, and *R* is surface reflectance for SWIR electromagnetic domain (1650 nm corresponding to band 6 of the Landsat 8).

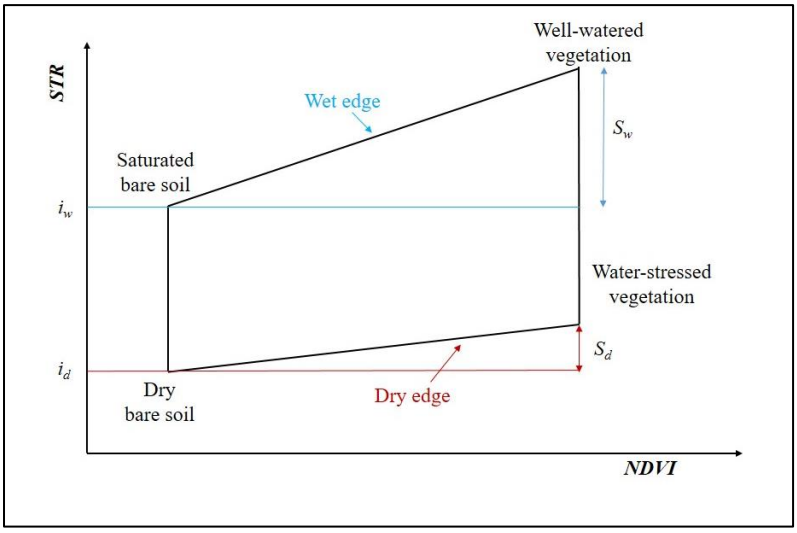

Figure 3.2. Sketch illustrating parameters of the optical trapezoidal model (OPTRAM) (Equation v). OPTRAM is parameterized based on the pixel distributions within the *STR-NDVI* space and was proposed by Sadeghi et al. (2017).

The *STR-NDVI* space forms a trapezoid shape based on the assumption of a linear relationship between soil- and vegetation-water contents. Therefore, the parameters for equation (3.1) can be obtained for a specific location in the satellite scene from the dry and wet edges of the optical trapezoid;

$$STR_d = i_d + S_d NDVI, \tag{3.3}$$

$$STR_w = i_w + S_w NDVI, \tag{3.4}$$

where, i_d and S_d are the intercept and slope of the dry edge and i_w and S_w are the intercept and slope of the wet edge. Based on the equations (3.1) - (3.4), soil moisture within a given satellite image pixel can be estimated from its STR and NDVI values:

$$W = \frac{i_d + S_d NDVI - STR}{i_d - i_w + (S_d - S_w) NDVI}, \tag{3.5}$$

where, the saturation degree W can be expressed as θ ($m^3\ m^{-3}$) when multiplied with the soil porosity.

Google earth engine (GEE) was used to process Landsat 8 images for the study period and area to obtain soil moisture maps via the OPTRAM model. GEE provides access to archived Landsat data, which includes Landsat 5, Landsat 7 and Landsat 8 OLI/TIRS from 2013 (Google Earth Engine, 2012; Google Earth Engine: A planetary-scale platform for Earth science data & analysis; URL: https://earthengine.google.com; accessed 01/06/2020). GEE was used to obtain STR, NDVI and OPTRAM surface soil moisture maps as suggested by Yadav et al. (2019) and Huang et al. (2017).

Standardized precipitation index (SPI)

The change in precipitation over time has direct impacts on groundwater, reservoir storage, surface soil moisture, snowpacks and stream flow. SPI is a statistical method for assessing rainfall developed by McKee et al (1993). It is often preferred over the mean

precipitation as a representation of what is the normal daily rainfall. For calculating SPI, the observed rainfall values from previous years are fitted to a gamma distribution and then transformed to a Gaussian distribution (Abramowitz and Stegun, 1948). For this study, we used R statistical software 2020 using the *precintcon* R-package.

SPI was designed to quantify the precipitation deficit for multiple timescales, which reflects the impact of drought based on the rainfall in specific year. In this study, we are establishing precipitation and surface soil moisture that responds on a relatively short time scale. Precipitation data were collected from the weather stations since their establishment date and the SPI was calculated. The SPI values range from ±2, where +2 is extremely wet, -2 is extremely dry, and -1 to +1 is near normal conditions (NDMC, 2008). For this study, we calculated monthly SPI values that displayed the percentage of normal precipitation for 30-day periods.

Model development and workflow

The following paragraphs describe the model development and workflow for objectives 2, 3, and 4 in this study (Figure 3.3). For objective 2, the Landsat 8 images for the study area through the 2019 growing season were processed by using ENVI software to establish relationships between the various vegetation indices described in section 2.3.1 above with the corresponding in-situ surface soil moisture. For objective 3, Google Earth Engine was used to prepare soil moisture saturation maps via the OPTRAM model proposed by Sadeghi et al. (2017) and used by Yadav et al. (2019) and Huang et al. (2017). Surface moisture for each location pixel was extracted from the OPTRAM moisture map using ArcGIS Pro® to establish relationship with in-situ surface soil moisture (Figure 3.3).

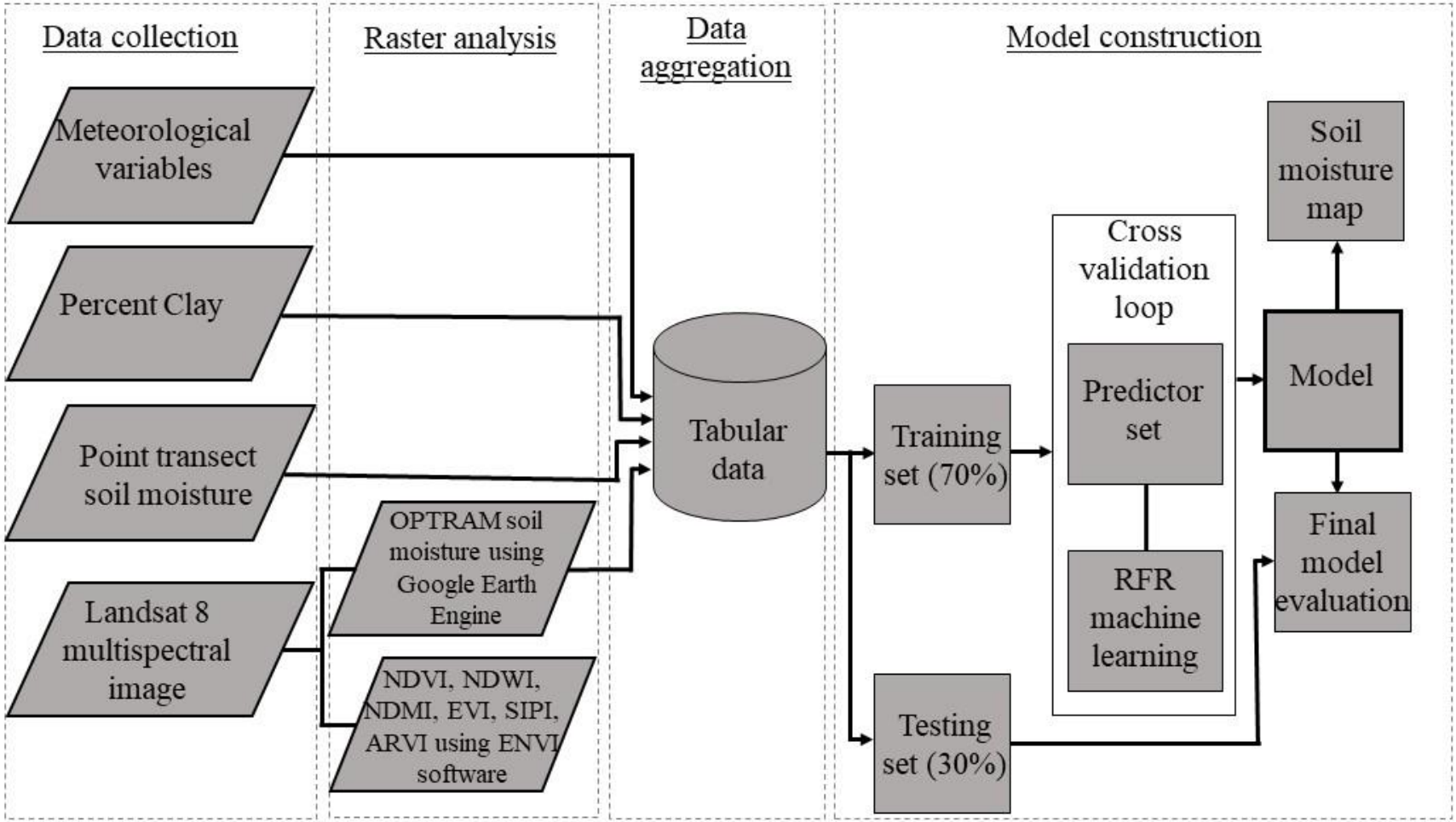

Figure 3.3. Flow chart showing the Landsat 8 image processing and soil moisture prediction model using Random forest regression (RFR) algorithm

For objective 4, rainfall data from 25 weather stations for entire 2019 growing season were downloaded from NDAWN website (https://ndawn.ndsu.nodak.edu) and four-day cumulative rainfall (mm) for each soil sampling dates (day of Landsat 8 satellite overhead pass) were calculated. Four-day cumulative rainfall maps were created using the rainfall values and ArcGIS Pro® ordinary kriging geospatial tool. SPI maps were also created using ordinary kriging for each month based on SPI values previously calculated in the R-software environment. The SPI maps were created for three months (June, July and August) (Figure 3.10) and cumulative rainfall maps were developed for four dates (June 18, July 20, August 5 and August 21) (Figure 3.11) for the 2019 growing season. Percent clay were extracted from the map downloaded from the University of California Davis – California soil resources lab and extracted for each sampling location. The pixel size of the maps was maintained at 30 m × 30 m.

The OPTRAM model's soil moisture, four-day cumulative rainfall, SPI and percent clay values were all extracted from respective maps for each soil sampling location. These four variables were used as predictor variables for in-situ surface soil moisture in a random forest regression model (RFR) as described in Breiman (2001). The whole dataset was divided into training and testing data at 70:30 ratio. Training dataset was used to develop the random forest regression algorithm and testing dataset was used for validation. The validation was assessed by the model's root mean square error (RMSE) and coefficient of determination (r^2).

The trained and validated random forest regression algorithm was then applied to selected agricultural fields within Cass County, North Dakota to estimate surface soil moisture. The OPTRAM soil moisture, percent clay, four-day cumulative rainfall and SPI was extracted from each pixel (30 m × 30 m) in the selected area. The extracted values were then used as input to the

trained random forest algorithm in the R-software environment to predict the area's surface soil moisture and the output mapped in ArcGIS Pro® using the pixel-by-pixel method.

Results

Relationship between vegetation indices and surface soil moisture

No relationship was observed between the vegetation indices and the surface soil moisture content for the entire study area and growing season (Figure 3.4). The scatter plot showed none of the sample points lie around the bisector line (red colored line) for all the vegetation indices, with r^2 values less than 0.10 and slopes less than 0.20. The NDVI values for corn increased with vegetative growth after the emergence until the silking stage (Figure 3.5). Afterwards, NDVI values remained the same until the crop started drying and was ready for harvest. The rate of increase is different in the soybean and wheat but NDVI values remain mostly constant after closure of the leaf canopies and crops grain fill. The NDVI values after planting remained the same for 80 days for all crops until drying and harvesting.

Early in the growing season, there was some bare ground on corn and soybean fields, but the wheat field was quickly covered with greenness after emergence. This can be observed in the NDMI values (Figure 3.5), where early stages in the corn and soybean crops showed negative values; however, there were positive values in the wheat fields. In NDMI values, variation was observed until the full canopy was observed in all crops, and not much change was observed afterwards. Unlike the NDVI and NDMI, the values are less for NDWI and approaches -1 when the crop is at peak vegetative growth. A similar trend to NDVI was also observed with EVI and ARVI for all crops and days after planting. SIPI for all crops did not change much with growth stages because it is based on the leaf area index, chlorophyll, and carotenoid content. The values for each crop are different, but changes within the crop growth stages are minimal.

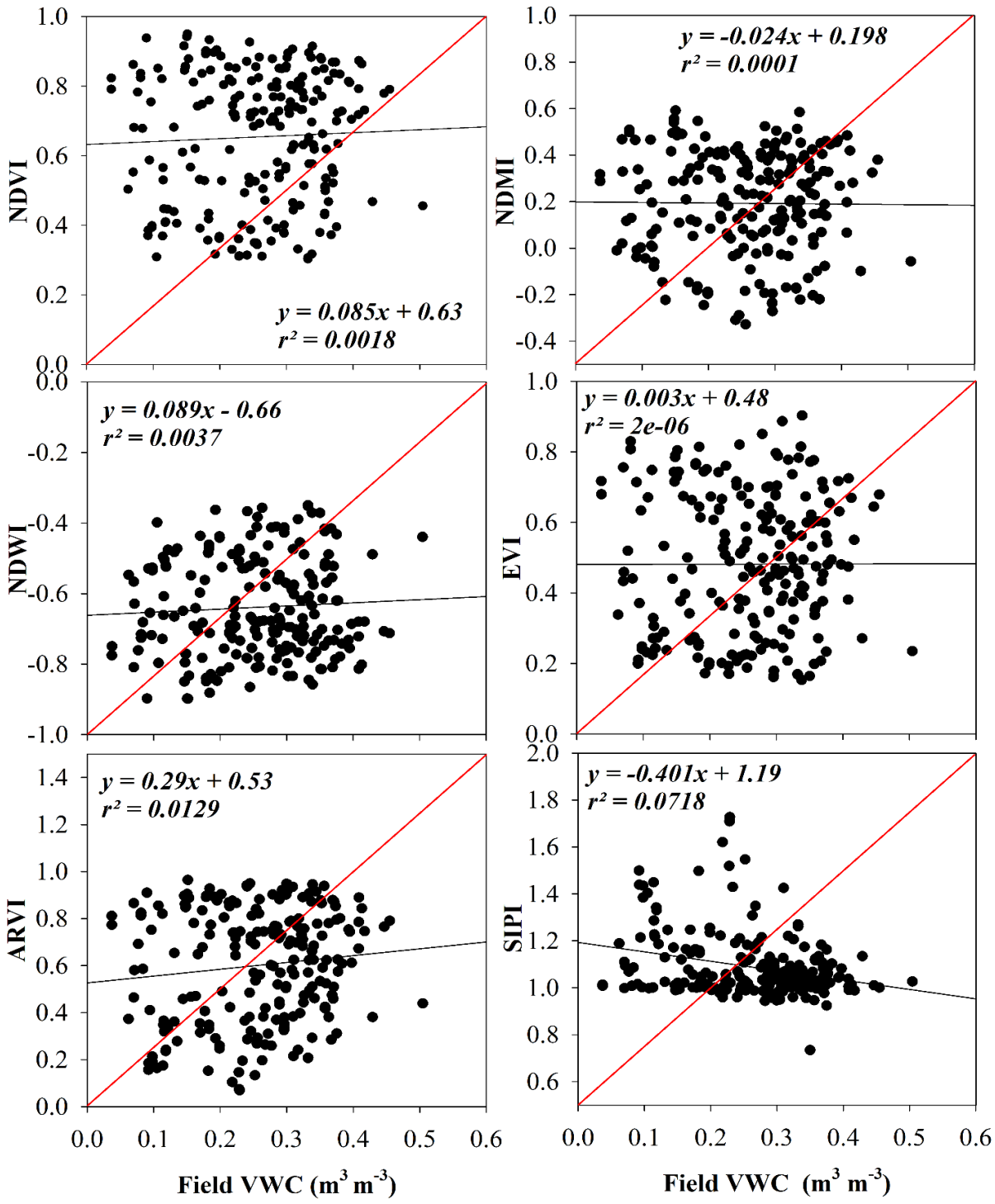

Figure 3.4. Scatter plot showing observed field VWC (m^3 m^{-3}) versus vegetation indices (NDVI, NDMI, NDWI, EVI, ARVI, SIPI) for 2019 growing season in the RRVN using Landsat 8 images along with regression coefficient (r^2) and linear equation.

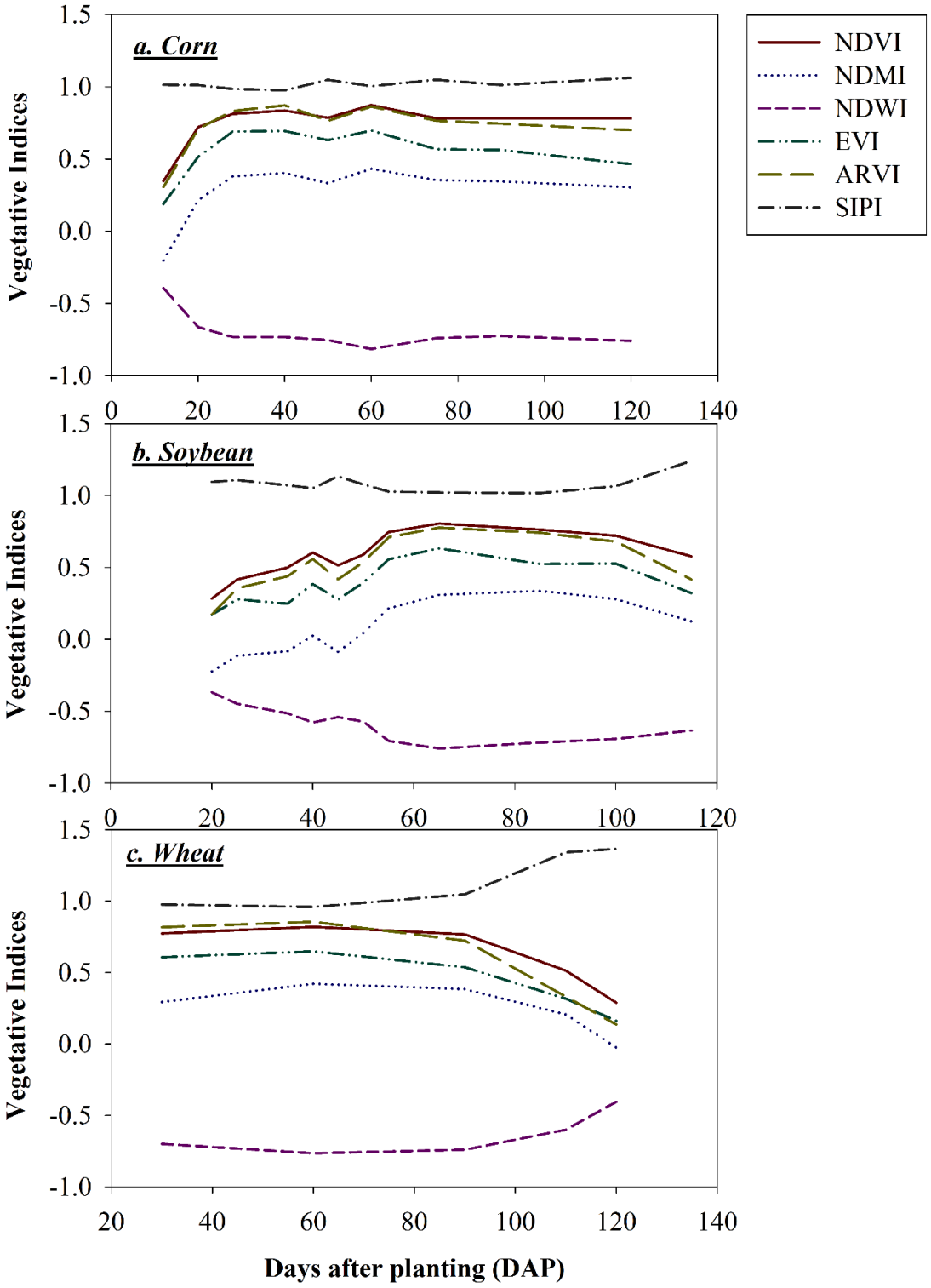

Figure 3.5. Changes in the values of vegetation indices (NDVI, NDMI, NDWI, EVI, ARVI, SIPI) with the days after planting for three crops (corn, soybean and wheat) for 2019 growing season in the RRVN.

Surface soil moisture prediction using OPTRAM

Pixel distribution within the STR-NDVI space for different days (6/18-8/21) did not show distinct trapezoidal shapes (Figure 3.6). Each of the six days exhibited a wide range of SRT and NDVI values due to variability of soil moisture and land cover. In general, STR and NDVI values are larger at the peak vegetative state (7/27/2019) as compared to early and later growth stages of crop.

Soil moisture maps created with OPTRAM model showed the surface soil moisture values were higher in the month of July and early August compared to early June and later August (Figure 3.7). However, the performance of OPTRAM with predicting the surface soil moisture was poor as shown in terms of the regression slope, r^2 and RMSE (Figure 3.8). The regression slopes for all dates are less than 0.20 with negative slope values on multiple dates.

Surface soil moisture mapping with a random forest algorithm

The RFR algorithm's validation showed a high goodness of fit (r^2=0.69) and low RMSE value (0.053 m^3 m^{-3}) (Figure 3.9) with just the four predictor variables (SPI, four-day cumulative rainfall, percent clay values, and the OPTRAM soil moisture). The SPI for each month (June, July and August) and four-day cumulative values for the specific dates of satellite passage (June 18, July 20, August 5 and 15) showed a wide range of values. The SPI values ranged from -0.26 to 0.94 for June, -0.78 to 1.40 for July and -0.14 to 1.40 for August, 2019 (Figure 3.10). This showed that July had normal precipitation over the study area, and July and August months were wetter this year compared to previous years. Negative values of the SPI indicate less than median precipitation (dry conditions) and positive values showed greater than median precipitation (wet conditions). However, four-day cumulative rainfall for June 18 was observed lowest for the entire study area (5.38 to 21.55 mm) compared to August 21, which ranged from 0 to 48.20 mm.

For July 20 and August 5, the cumulative rainfall was similar ranging from 0 to 37.37 mm and 0.31 to 37.22 mm (Figure 3.11). Some parts of the study area have not received rainfall for past four days in two dates (July 20 and August 21).

Each pixel values of SPI, four-day cumulative rainfall, percent clay, and OPTRAM soil moisture among the select are of agricultural fields in Cass County, ND were used to demonstrate mapping of the algorithm's surface soil moisture (Figure 3.12). The surface soil moisture for all four dates ranges from 0.22 to 0.39 m^3 m^{-3}. The surface soil moisture values on June 18 and August 21 were less compared to July 20 and August 3.

Discussion

Surface soil moisture estimation using vegetation indices

Vegetation indices used in this study are interrelated based on the Landsat 8 bands used to calculate them (Figure 3.13). This interrelation is likely part of the reason why all the vegetation indices were similar in their poor performance for indicating surface soil moisture. The NIR band is common in the calculation of all vegetation indices. In addition to NIR band, NDVI used red band, NDWI used green band, NDMI used SWIR band, SIPI used blue band for calculation. EVI and ARVI are two indices that have used red and blue bands along with NIR. Green vegetation strongly reflects NIR bands and green vegetation is directly related to the moisture of soil where it is grown. Red band of the spectrum is absorbed by the vegetation, SWIR band discriminates moisture content of soil and vegetation. Blue band is sensitive to the chlorophyll and carotenoid molecules of the plant leaf, and the green band boosts water information. All the vegetation indices are useful tools to measure greenness based on the crop vegetative growth, but are not useful to indicating surface soil moisture in the RRVN.

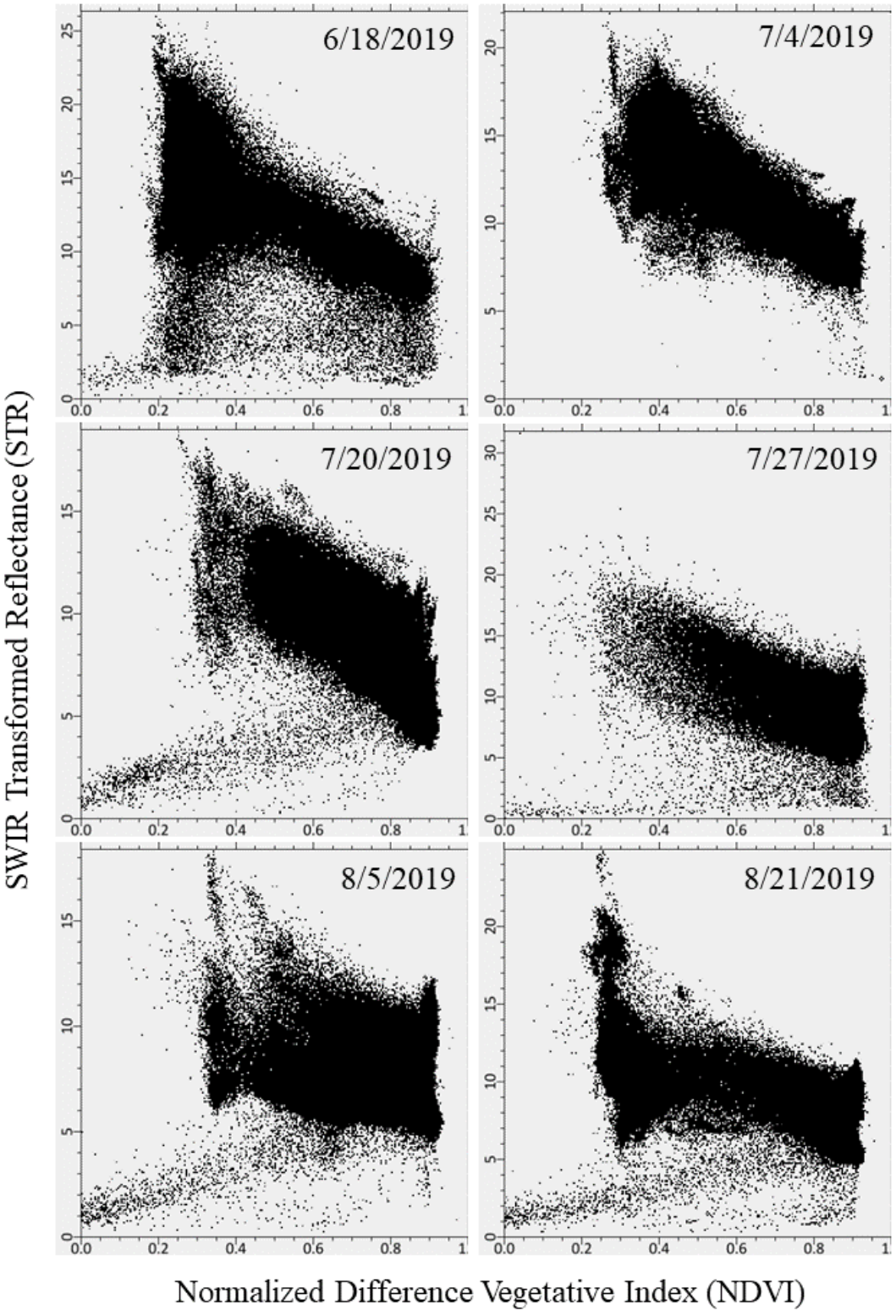

Figure 3.6. Pixel distributions within the SRT-NDVI space for all the image for 2019 growing season (6/18 to 8/21) in the RRVN.

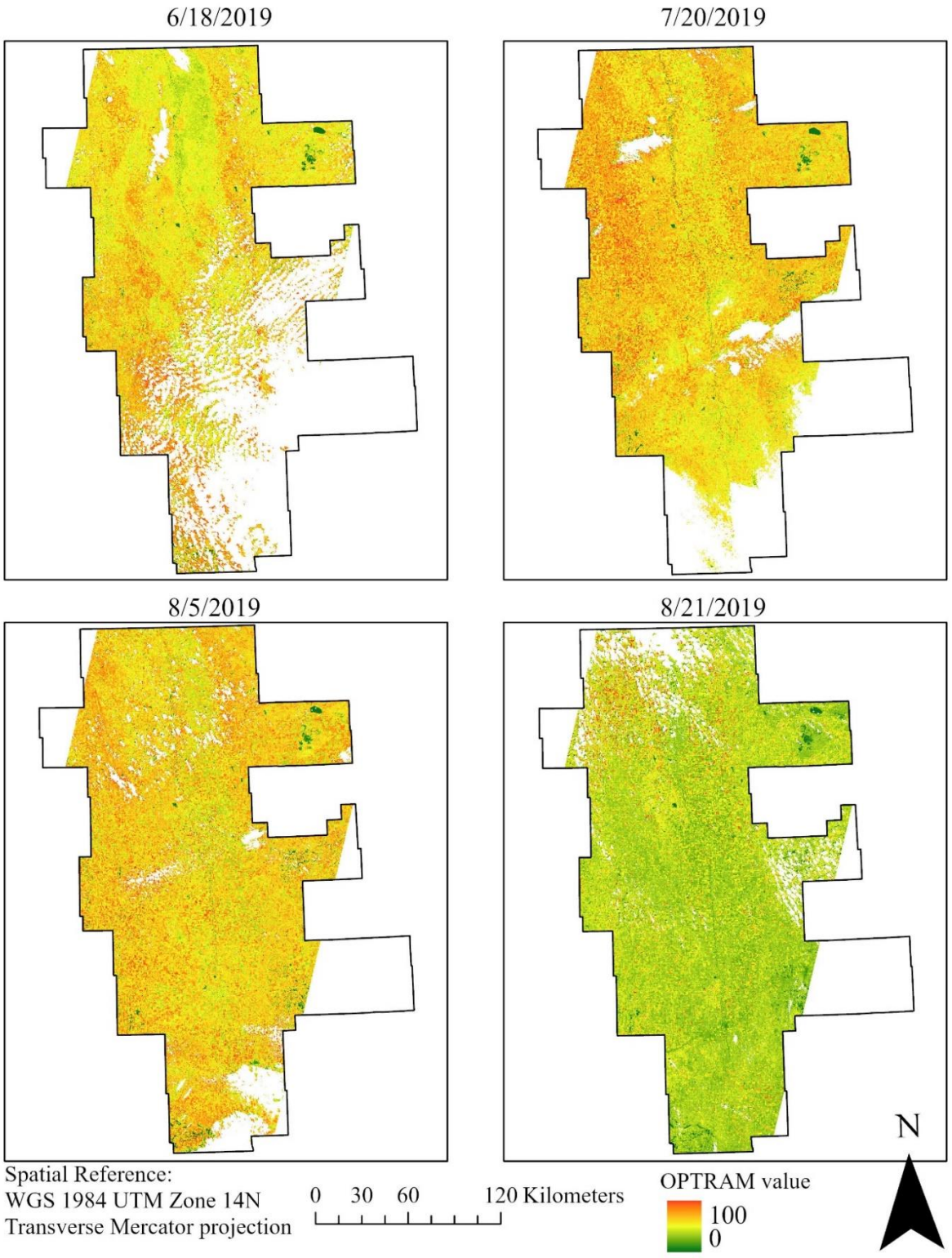

Figure 3.7. Surface soil moisture maps generated with OPTRAM using Landsat 8 images for different dates (6/18, 7/20, 8/5, 8/21) of 2019 growing season in the RRVN. White pixels represent maxed pixel due to water bodies, shadows, clouds, and rural/urban areas.

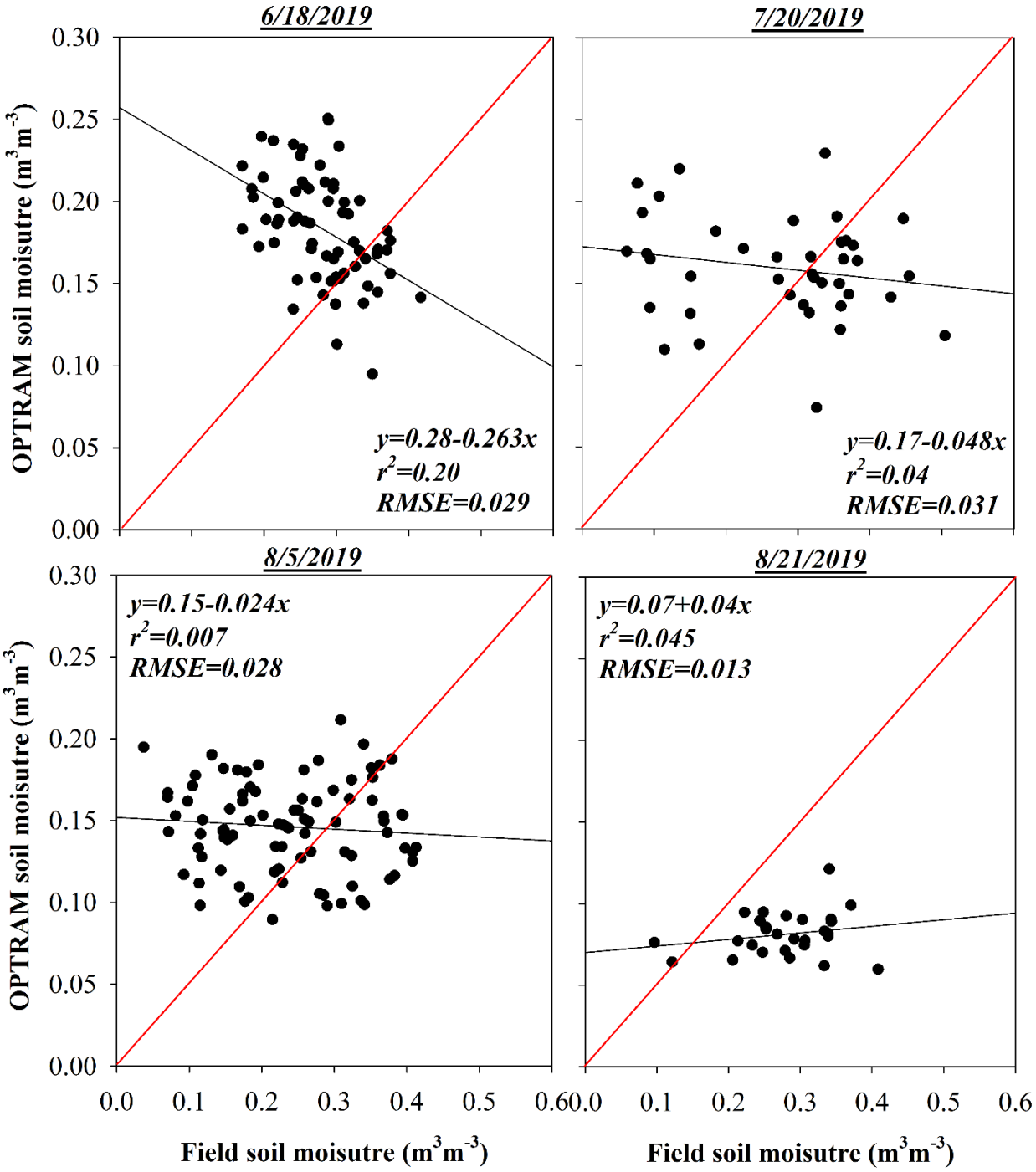

Figure 3.8. Soil moisture estimated by OPTRAM compared to field soil moisture for different dates during 2019 growing season in the RRVN.

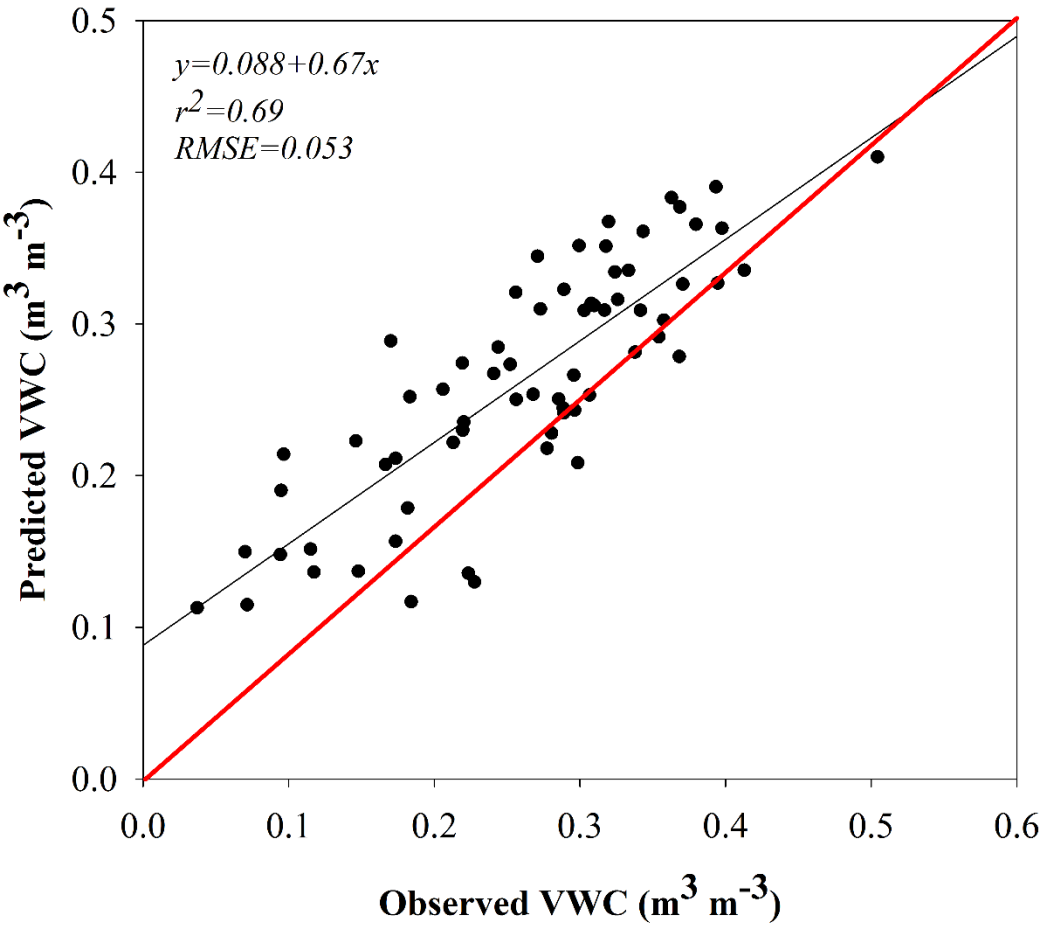

Figure 3.9. Scatter plot showing observed versus predicted volumetric water content (m^3 m^{-3}) during the testing phase along with regression coefficient (r^2) and root mean square error (RMSE) for random forest regression model.

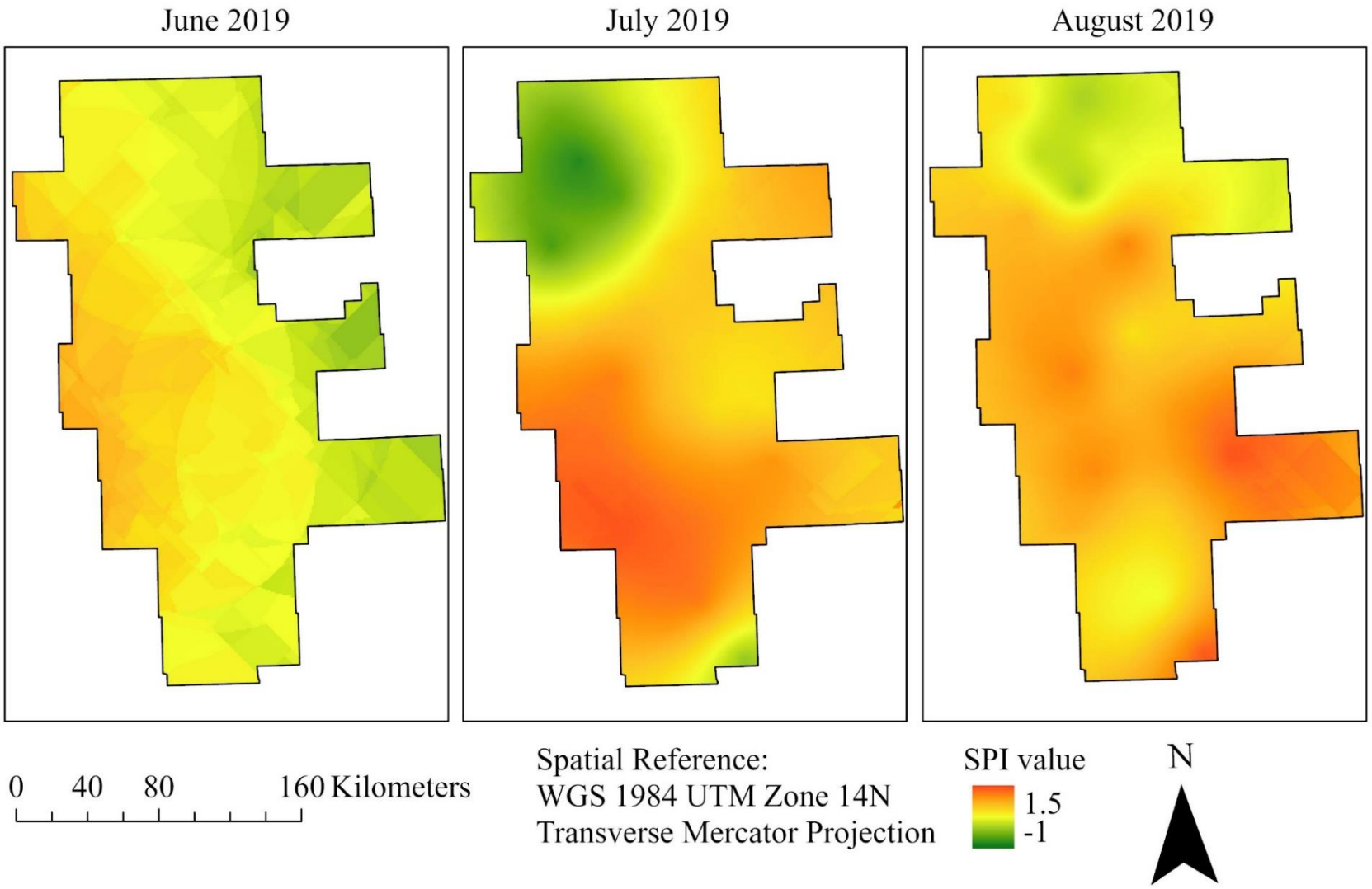

Figure 3.10. Standardized Precipitation Index (SPI) maps created using rainfall data from 25 weather stations by ordinary kriging interpolation in the RRVN for June, July and August month of 2019.

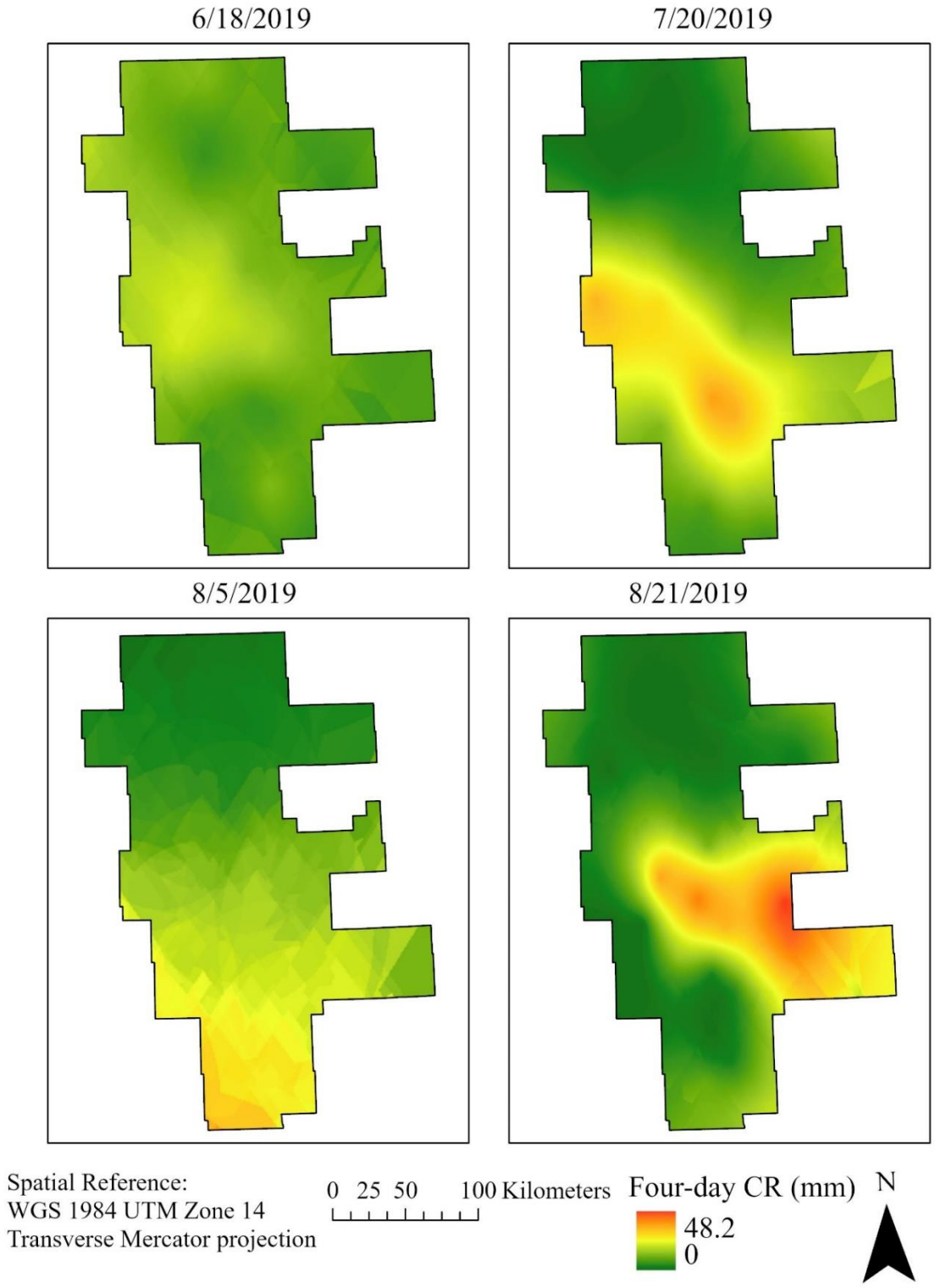

Figure 3.11. Four-days cumulative rainfall maps created using past four-days rainfall data from 25 weather stations by ordinary kriging interpolation in the RRVN for 2019 growing season (6/18, 7/20, 8/5 and 8/21).

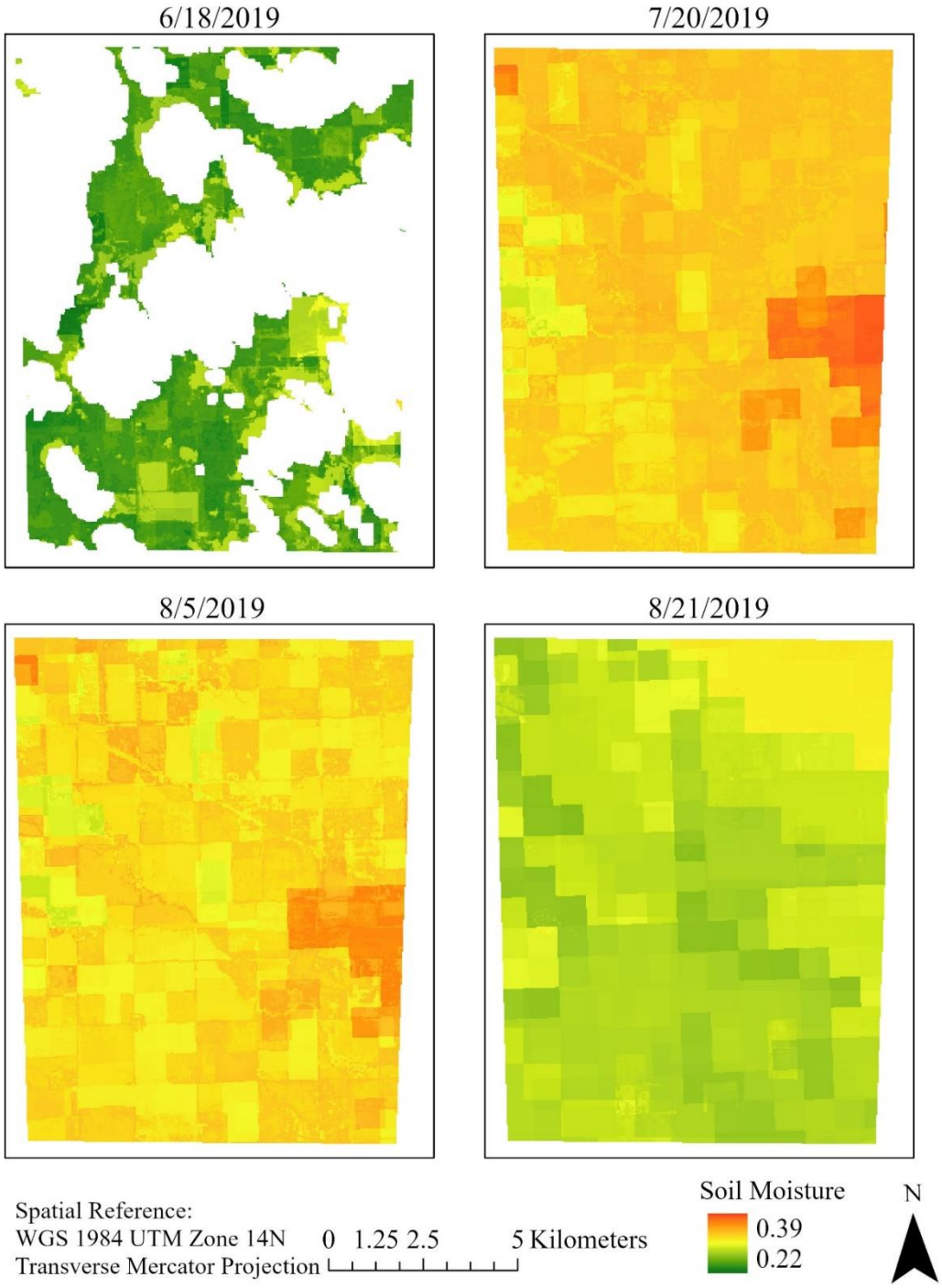

Figure 3.12. Predicted surface soil moisture ($m^3\ m^{-3}$) for four days (6/18, 7/20, 8/5 and 8/21) over selected agriculture field in Cass County within the study area.

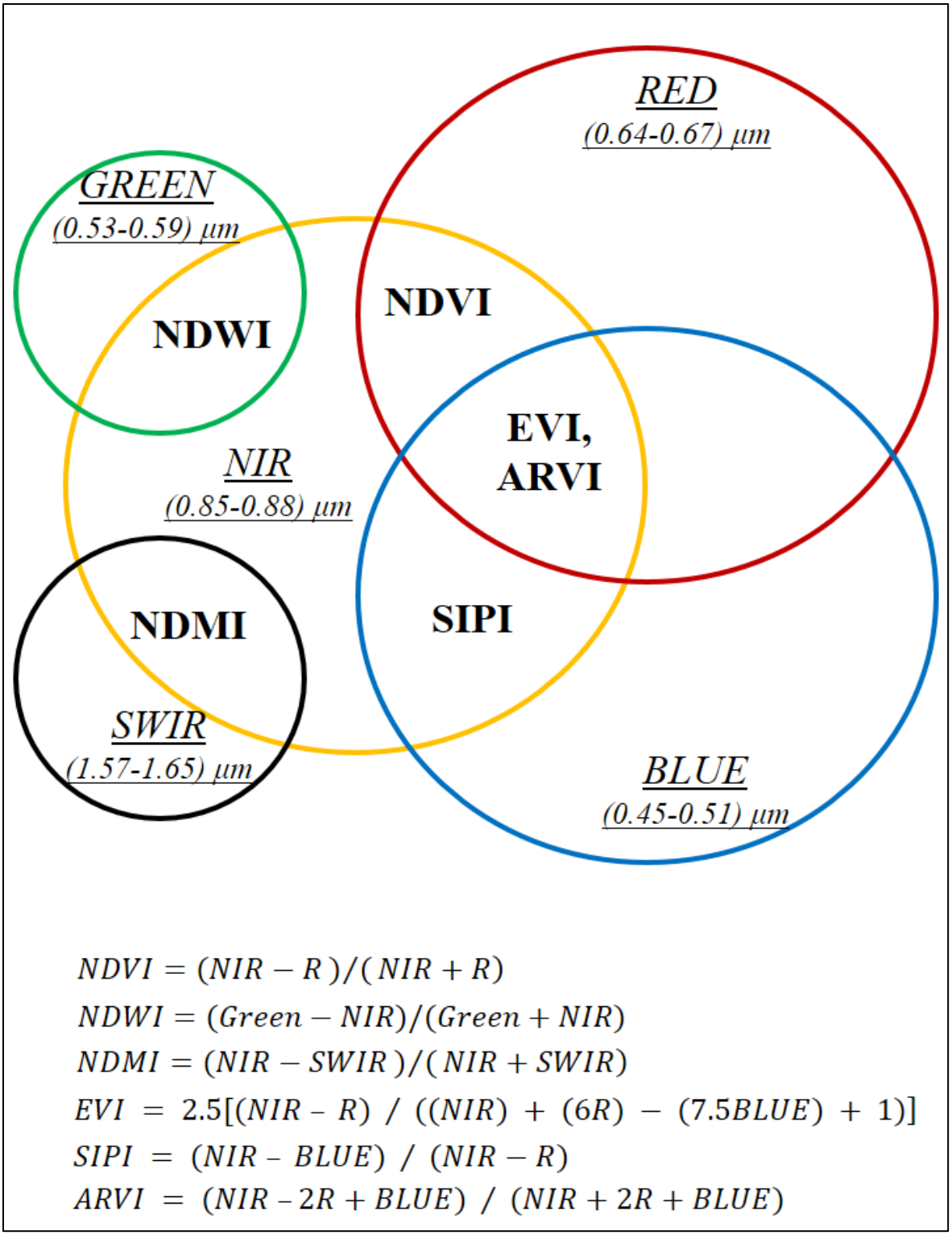

Figure 3.13. Relationship between the vegetation indices on how they are calculated using different Landsat 8 image bands (red, green, blue, near infrared, short wave infrared) and their formula.

Cushion et al. (2005) also found little to no relationship between NDVI and surface soil moisture in time series data. In contrast, Martyniak et al. (2007) claimed that water availability in the root zone is one of the main factors controlling crop growth and the vegetation status, hence soil moisture can be represented by the NDVI. Wang et al. (2007) concluded mapping root-zone soil moisture is challenging because the relationship between NDVI and root-zone soil moisture is dependent on the vegetation species and climate zones. They noted NDVI derived from satellite images may provide a proxy for root-zone soil moisture mapped at large scales. Adegoke and Carleton (2002) found neutron probe measurements of soil moisture taken in forest and crop sites are weakly correlated with full-resolution NDVI for each pixel. They also found the association measured by Pearson correlation coefficient between NDVI and soil moisture is stronger over forest than over the cropland during the growing season. NDVI along with other vegetation indices (NDMI, NDWI, EVI, ARVI and SIPI) showed the same trend during the growing season for the three major crops in the RRNV of our study. The region of the RRVN in this study has a continental climate (Bell and Halpert, 1998), thick underlying glaciolacustrine deposits of low permeability, and shallow perched water tables (Remenda et al., 1994). We observed that fields did not dry to levels where plants were water stressed enough to affect crop physiology and vegetative growth. This may be due to adequate root-zone soil moisture and shallow ground water tables serving as a supply of water even when surface soil moisture is quiet low. This would likely result in the low variability in the values of vegetation indices.

Effectiveness of OPTRAM to predict surface soil moisture

The weak relationship between the OPTRAM model's estimated soil moisture and the actual field surface soil moisture may have been partly due to high values of SRT coupled with low NDVI in peak vegetative stage. This can occur when there is water in excess of saturated

soil moisture. The soils in the RRVN are derived from an ancient glaciolacustrine lakebed with high clay content, poor internal drainage, and are prone to flooding. Water in excess of the saturated soil moisture will cause an increase in STR while the actual soil moisture cannot increase beyond the saturated soil moisture content (Sadeghi et al., 2015, 2017).

Similar results were observed by Babaeian et al. (2018) in across the diverse climate of the United States, where r^2 ranges from 0.01 to 0.49 and RMSE ranges from 5-8%. They used the OPTRAM model to estimate soil moisture in Arizona, California and Georgia, and found nearly trapezoidal shape by STR-NDVI points. They observed that distributions of STR-NDVI points differ with the area under study due to characteristics of climate and land cover. Yadav et al. (2019) also used OPTRAM model to predict soil moisture in Lalitpur district of India and found the accuracy of this model was less than 10% when compared with other models (thermal). However, Yadav et al. (2019) concluded this reduced accuracy might be due to lack of penetration of optical bands in the fully or densely covered vegetation. Similarly, Chen et al. (2020) evaluated OPTRAM-based soil moisture estimates using MODIS data provide overall RMSE from 0.05 to 0.13 $m^3 m^{-3}$, and r^2 from 0.10 to 0.50 respectively, which corroborate with our results. The poor performance between the OPTRAM soil moisture estimates and in situ soil moisture might be due to rough image resolution and heterogeneous terrains (Chen et al., 2020).

Machine learning for soil moisture prediction

As anticipated, the RFR algorithm we developed substantially outperformed the other methods for predicting soil moisture in the RRVN. Moreover, this was accomplished with very few input variables (n=4) from typical weather stations and soil maps. Recently, machine learning algorithms for remote sensing data have been used elsewhere to predict soil moisture with variable success (Adab et al., 2020; Araya et al., 2020; Li et al., 2020; Satalino et al., 2002;

Paloscia et al., 2013). Adab et al. (2020) compared four different machine learning models (support vector machine (SVM), artificial neural network (ANN), elastic net regression, and RFR) for predicting near-surface soil moisture in semi-arid Iran using Landsat-8 images and found highest explanatory ability (NS = 0.73) with RFR algorithm. Similarly, Araya et al. (2020) compared five machine learning models (ANN, SVM, support vector regression, relevance vector regression, and boosted regression trees) using unmanned aircraft system (UAS) to capture reflectance (green, red and near infrared) along with topographic and meteoric (rainfall and precipitation) variables to predict surface soil moisture. They found RFR and BRT models performed better with error less than 4% volumetric soil water content.

Overall, the integration of machine learning (particularly RFR) with physically-based remote-sensing models and commonly available weather and soil data appears to be a reliable tool for estimating and mapping surface soil moisture at high-resolutions across temperate, arid, and now frigid soil landscapes, deep and shallow water tables, and with a wide range of homogenous and heterogeneous vegetation. Future research efforts should be aimed at directly resolving the STR-NDVI space issues of the OPTRAM model in landscapes with frequent or prolonged flooding, discontinuous water-filled potholes, and permafrost and/or aimed at indirectly resolving them by identifying key parameters to include in the machine learning algorithm input.

Conclusion

Remote sensing and machine learning are powerful tools that not only can cover a large area but also process a large amount of dataset and can be used in moisture prediction. This study showed the integrated use of satellite images, weather stations and soil properties to predict surface soil moisture of agricultural fields. Six vegetation indices calculated by using ENVI

platform showed poor relationship with infield soil moisture. The weak relationship was established due to the difference in growth stages of crop and canopy cover. In extreme drought conditions vegetation indices may reflect low soil moisture content. OPTRAM soil moisture maps developed by using google earth engine also showed weak relationship with infield surface moisture content. However, the soil moisture prediction improved significantly after OPTRAM values were incorporated with rainfall, SPI and percent clay using random forest regression machine learning algorithm. This research proposed an integrated model to predict surface soil moisture over a larger area using satellite images and weather stations that are easily available. However, this model has to be tested in different areas and consider other possible factors that can affect surface soil moisture depending on the landscape and location.

References

Abramowitz, M., Stegun, I.A., 1948. Handbook of mathematical functions with formulas, graphs, and mathematical tables. Vol. 55. US Government printing office.

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., Ghalhari, G.A.F., 2020. Machine learning to estimate surface soil moisture from remote sensing data. Water, 12(11), 3223.

Adegoke, J.O., Carleton, A.M., 2002. Relations between soil moisture and satellite vegetation indices in the US Corn Belt. J Hydrometeorol, *3*(4), 395-405.

Alfieri, L., Claps, P., D'Odorico, P., Laio, F., Over, T.M., 2008. An analysis of the soil moisture feedback on convective and stratiform precipitation. J. Hydrometeorol., 9, 280–291, https://doi.org/10.1175/2007jhm863.1.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M. and Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens., 7(12), 16398-16421.

Araya, S.N., Fryjoff-Hung, A., Anderson, A., Viers, J.H., Ghezzehei, T.A., 2020. Advances in soil moisture retrieval from multispectral remote sensing using unmanned aircraft systems and machine learning techniques. Hydrol. Earth Syst. Sc., 1-33.

Babaeian, E., Sadeghi, M., Franz, T.E., Jones, S., Tuller, M., 2018. Mapping soil moisture with the OPtical TRApezoid Model (OPTRAM) based on long-term MODIS observations. Remote Sens. Environ., 211, 425-440.

Bastiaanssen, W.G., Molden, D.J., Makin, I.W., 2000. Remote sensing for irrigated agriculture: Examples from research and possible applications. Agr. Water Manage. 46, 137-155.

Bell, G.D., Halpert, M.S., 1998. Climate assessment for 1997. Bulletin of the American Meteorological Society, 79(5s), S1-S50.

Breiman, L., 2001. Random Forest, Mach. Learn. 45(1), 5–32, doi:10.1023/A:1010933404324.

Brocca, L., Moramarco, T., Melone, F., Wagner, W., 2013. A new method for rainfall estimation through soil moisture observations. Geophys. Res. Lett., 40(5), 853-858.

Cashion, J., Lakshmi, V., Bosch, D., Jackson, T.J., 2005. Microwave remote sensing of soil moisture: evaluation of the TRMM microwave imager (TMI) satellite for the Little River Watershed Tifton, Georgia. J. Hydrol., 307(1-4), 242-253.

Ceccato, P., Flasse, S., Tarantola, S., Jacquemond, S., Gregoire, J., 2001. Detecting vegetation water content using reflectance in the optical domain. Remote Sens. Environ., 77, 22–33.

Das, K., Paul, P.K., 2015. Present status of soil moisture estimation by microwave remote sensing. Cogent Geoscience, 1(1), 1084669.

Dupigny-Giroux, L.A., Lewis, J.E., 1999. A moisture index for surface characterization over a semiarid area. Photogramm. Eng. Rem. S., 65, 937–945.

Francois, C., 2002. The potential of directional radiometric temperatures for monitoring soil and leaf temperature and soil moisture status. Remote Sens. Environ., 80, 122–133.

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

Hu, X., Xue, M., McPherson, R.A., 2017. The importance of soil-type contrast in modulating August precipitation distribution new the Edwards Plateau and Balcones Escarpment in Texas. J. Geophys. Res-Atmos., 122(10), 711-728, https://doi.org/10.1002/2017JD027035.

Huang, H., Chen, Y., Clinton, N., Wang, J., Wang, X., Liu, C., Gong, P., Yang, J., Bai, Y., Zheng, Y., Zhu, Z., 2017. Mapping major land cover dynamics in Beijing using all Landsat images in Google Earth Engine. Remote Sens. Environ., 202, 166-176

Jackson, T.J., Chen, D., Cosh, M., Li, F., Anderson, M., Walthall, C., Doriaswamy, P., Hunt, E.R., 2004. Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans. Remote Sens. Environ., 92(4), 475-482.

Kaufman, Y.J., Tanre, D., 1992. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. IEEE T. Geosci. Remote, 30(2), 261-270.

Leng, P., Li, Z., Duan, S., Gao, M., Huo, H., 2017. A practical approach for deriving all-weather soil moisture content using combined satellite and meteorological data. ISPRS J. Photogramm., 131, 40-45. doi: 10.1016/j.isprsjprs.2017.07.013.

Leng, P., Song, X., Duan, S., Li, Z., 2016. A practical algorithm for estimating surface soil moisture using combined optical and thermal infrared data. Int. J. Appl. Earth. Obs., 52, 338-348. doi: 10.1016/j.jag.2016.07.004.

Levit, D.G., Simpson, J.R., Huete, A.R., 1990. Estimates of surface soil water content using linear combinations of spectral wavebands. Theor. Appl. Climatol., 42, 245–252.

Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Sci. Rep., 10(1), 1-12.

Liang, S.L., 2004. Quantitative remote sensing of land surface. John Wiley & Sons: Hoboken, NJ, USA, 2004.

Lillesand, T.M., Kiefer, R.W., Chipman, J.W., 2008. Remote sensing and image interpretation. Hoboken, NJ: John Wiley & Sons.

Liu, H.Q., Huete, A., 1995. A feedback-based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE T. Geosci. Remote, 33(2), 457-465.

Martyniak, L., Dabrowska-Zielinska, K., Szymczyk, R., Gruszczynska, M., 2007. Validation of satellite-derived soil-vegetation indices for prognosis of spring cereals yield reduction under drought conditions–Case study from central-western Poland. Adv. Space Res., 39(1), 67-72.

Mcfeeters, S.K., 1996. The use of Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens., 17, 1425–1432.

Mckee, T.B.N., Doesken, J., Kleist, J., 1993. The relationship of drought frequency and duration to time scales. Proceedings of the Eighth Conference on Applied Climatology, Anaheim, CA (Anaheim, CA: American Meteorological Society). 179–184.

Mckee, T.B.N., Doesken, J., Kleist, J., 1995. Drought monitoring with multiple time scales. Proceedings of the Ninth Conference on Applied Climatology, Dallas, TX (Dallas, TX: American Meteorological Society). 233–236.

NDAWN, 2020. North Dakota Agricultural Weather Station. https://ndawn.ndsu.nodak.edu/ (data retrieved on 5/8/2020).

NDMC, 2008. National Drought Mitigation Center. Explanation of the US drought monitor. http://droughtmonitor.unl.edu/classify.htm. Accessed 15 July 2020.

Paloscia, S., Pampaloni, P., Pettinato, S., Santi, E., 2008. A comparison of algorithms for retrieving soil moisture from ENVISAT/ASAR images. IEEE T Geosci. Remote, 46(10), 3274-3284.

Paloscia, S., Pettinato, S., Santi, E., Notarnicola, C., Pasolli, L., Reppucci, A.J.R.S.O.E., 2013. Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation. Remote Sens. Environ., 134, 234-248.

Peng, J., Loew, A., Merlin, O., Verhoest, N.E., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys., 55(2), 341–366. doi:10.1002/2016RG000543.

Penuelas, J., Baret, F., Filella, I., 1995. Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance. Photosynthetica, 31(2), 221-230.

Penuelas, J., Filella, I., Biel, C., Serrano, L., Save, R., 1993. The reflectance at the 950– 970 mm region as an indicator of plant water status. Int. J. Remote Sens., 14, 1887-1905.

Rahimzadeh-Bajgiran, P., Berg, A.A., Champagne, C., Omasa, K., 2013. Estimation of soil moisture using optical/thermal infrared remote sensing in the Canadian Prairies. ISPRS J Photogramm., 83, 94-103.

Remenda, V.H., Cherry, J.A., Edwards, T.W.D., 1994. Isotopic composition of old ground water from Lake Agassiz: Implications for late Pleistocene climate. Science, 266(5193), 1975-1978.

Reynolds, S.G., 1970. The gravimetric method of soil moisture determination Part I: A study of equipment, and methodological problems. J. Hydrol., 11(3), 258-273.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS. Third Earth Resources Technology Satellite-1 Symposium. Technical presentations, section A, Vol I: 309 – 317). Washington, DC: National Aeronautics and Space Administration (NASA SP-351).

Sadeghi, A.M., Jones, S.B., Philpot, W.D., 2015. A linear physically-based model for remote sensing of soil moisture using short wave infrared bands. Remote Sens. Environ., 164, 66–76. https://doi.org/10.1016/j.rse.2015.04.007.

Sadeghi, M., Babaeian, E., Tuller, M., Jones, S.B., 2017. The optical trapezoid model: a novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations. Remote Sens. Environ., 198, 52–68. https://doi.org/10.1016/j.rse.2017.05.041.

Sandholt, I., Rasmussen, K., Andersen, J., 2002. A simple interpretation of the surface temperature/vegetation index space for assessment of soil moisture status. Remote Sens. Environ., 79, 213–224.

Satalino, G., Mattia, F., Davidson, M.W., Le Toan, T., Pasquariello, G., Borgeaud, M., 2002. On current limits of soil moisture retrieval from ERS-SAR data. IEEE T. Geosci. Remote, 40(11), 2438-2447.

Shafian, S., Maas, S.J., 2015. Index of soil moisture using raw Landsat image digital count data in Texas high plains. Remote Sens., 7(3), 2352-2372.

Shi, L., Mao, Z., Chen, P., Gong, F., Zhu, Q., 2016. Comparison and evaluation of atmospheric correction algorithms of QUAC, DOS, and FLAASH for HICO hyperspectral imagery.

In Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions. 9999, 999917. International Society for Optics and Photonics.

Sims, A.P., Nigoyiand, D.S., Raman, S., 2002. Adopting indices for estimating soil moisture: a North Carolina case study. Geophys. Res. Lett., 29, Art. no. 1183.

Sims, D.A., Gamon, J.A., 2003. Estimation of vegetation water content and photosynthetic tissue area from spectral reflectance: a comparison of indices based on liquid water and chlorophyll absorption features. Remote Sens. Environ., 84, 526–537.

Sobrino, J.A., Raissouni, N., 2000. Toward remote sensing methods for land cover dynamics monitoring, application to Morocco. Int. J. Remote Sens., 20, 353-366.

Sun, H., 2016. Two-stage trapezoid: a new interpretation of the land surface temperature and fractional vegetation coverage space. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 9 (1), 336-346.

Szalai, S., Szinell, C., Zoboki, J., 2000. Drought monitoring in Hungary. In early warning systems for drought preparedness and drought management (Lisbon: World Meteorological Organization).182–199.

Tadesse, T., Brown, J., Hayes, M., 2005. A new approach for predicting drought-related vegetation stress: Integrating satellite, climate, and biophysical data over the U.S. central plains. ISPRS J. Photogramm., 59, 244-253

Tucker, C. J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ., 8, 127-150.

Tucker, C.J., 1980. Remote sensing of leaf water content in the near infrared. Remote Sens. Environ., 10, 23-32.

USDA, 2020. United States Department of Agriculture, International Production Assessment Division. Metadata for crops at different growth stage. https://ipad.fas.usda.gov/cropexplorer/description.aspx?legendid=312 (data retrieved on 5/8/2020)

Verstraeten, W.W., Veroustraete, F., Feyen, J., 2008. Assessment of evapotranspiration and soil moisture content across different scales of observation. Sensors 8(1), 70–117. doi:10.3390/s8010070.

Vicente-Serrano, S.M., Pons-Fernández, X., Cuadrat-Prats, J.M., 2004. Mapping soil moisture in the central Ebro river valley (northeast Spain) with Landsat and NOAA satellite imagery: a comparison with meteorological data. Int. J. Remote Sens., 25(20), 4325-4350.

Wang, L., Qu, J.J., 2009. Satellite remote sensing applications for surface soil moisture monitoring: A review. Front. Earth Sci-Prc. 3(2), 237-247. doi:10.1007/s11707-009-0023-7.

Wang, X., Xie, H., Guan, H. and Zhou, X., 2007. Different responses of MODIS-derived NDVI to root-zone soil moisture in semi-arid and humid regions. J. Hydrol., 340(1-2), 12-24.

West, H., Quinn, N., Horswell, M., White, P., 2018. Assessing vegetation response to soil moisture fluctuation under extreme drought using Sentinel-2. Water, 10(7), 838.

Yadav, S.K., Singh, P., Jadaun, S.P.S., Kumar, N. and Upadhyay, R.K., 2019. Soil moisture analysis of Lalitpur district Uttar Pradesh India using Landsat and sentinel data. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., Volume XLII-3/W6, ISPRS-GEOGLAM-ISRS Joint Int. Workshop on “Earth Observations for Agricultural Monitoring”, 18–20 February 2019, New Delhi, India

Yu, W., Ma, M., Li, Z., Tan, J., Wu, A., 2017. New Scheme for validating remote-sensing land surface temperature products with station observations. Remote Sens., 9(12), 1210.

Zeng, W., Xu, C., Huang, J., Wu, J., Tuller, M., 2016. Predicting near-surface soil moisture content of saline soils from NIR reflectance spectra with a Modified Gaussian model. Soil Sci. Soc. Am. J., 80, 1496-1506. http://dx.doi.org/10.2136/sssaj2016.06.0188.

Zhang, D., Tang, R., Zhao, W., Tang, B., Wu, H., Shao, K., Li, Z.L., 2014. Surface soil water content estimation from thermal remote sensing based on the temporal variation of land surface temperature. Remote Sens., 6, 3170-3187.

Zhang, D., Zhou, G., 2016. Estimation of soil moisture from optical and thermal remote sensing: a review. Sensors 16(8), 1308

Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsat 4–7, 8, and Sentinel 2 images. Remote Sens. Environ., 159, 269-277. http://dx.doi.org/10.1016/j.rse.2014.12.014.

GENERAL CONCLUSION

This study offers evidence that soil moisture can be reasonably represented by using information obtained at nearby weather stations despite large differences in soil and crop characteristics. The correlation between the soil moisture at weather stations and nearby agricultural fields is affected by crop type and their growth stages, crop residue, soil texture, and distance from the weather station. Similar associations were observed when crop growth stages were at peak vegetative and reproductive stages. However, higher correlations were observed with lower crop residue cover of the soil surface and vice-versa. Rainfall and evapotranspiration measured at weather stations can be used to estimate soil moisture in these nearby agricultural fields. The four-day cumulative rainfall and PET showed higher correlations with field soil moisture as compared to other durations. This shows that rainfall and precipitation data can be effectively used in the prediction on soil moisture in the nearby fields despite discrepancies in soil and crop characteristics. Machine learning algorithms can be used effectively in predicting field soil moisture. The RFR, BRT and SVR predictions performed better based on high correlations, low RMSE and MAE, during model validation using an independently derived dataset. The weather station variables (station soil moisture, four-day cumulative rainfall, and PET) were relatively more influential than the soil and crop variables for predicting field soil moisture in the nearby plots.

Remote sensing and machine learning are powerful tools that not only can cover a large area but also process a large amount of dataset and are used in moisture prediction. OPTRAM soil moisture maps developed by using google earth engine showed weak relationship with infield surface moisture content. However, the soil moisture prediction improved significantly after OPTRAM values were incorporated with rainfall, SPI and percent clay using random forest

regression machine learning algorithm. This research proposed an integrated model to predict surface soil moisture over a larger area using satellite images and weather stations that are easily available. Future research should evaluate the model in different regions and ecosystems where other landscape factors may have prominent effects on surface soil moisture.

APPENDIX

Table A1. Linear relationship between volumetric water content (VWC) of crop fields with nearby weather stations for different crop residue cover, crop type, distance from station and soil texture in study area.

	Intercept (c)	Gradient (m)	r^2	RMSE	N
		Overall			
	0.06692**	0.70216**	0.4977	0.06546	675
		Residue (%)			
>10	0.06437**	0.78619**	0.6255	0.05806	279
20-30	0.06475**	0.65107**	0.4434	0.06691	198
50-60	0.04933**	0.72622**	0.4612	0.06671	198
		Crop type			
Alfalfa	0.147*	0.45*	0.44	0.055	12
Barley	0.170	0.34	0.18	0.052	9
Canola	0.140**	0.71**	0.78	0.034	9
Corn	0.039	0.81**	0.46	0.068	156
Dey bean	0.066**	0.62**	0.69	0.041	33
Oats	0.012	0.68**	0.83	0.030	15
Potato	0.032	0.58*	0.83	0.022	6
Soybean	0.066**	0.69**	0.43	0.069	222
Sugar beet	0.064*	0.70**	0.54	0.066	45
Sunflower	-0.013	0.99	0.41	0.061	9
Wheat	0.076**	0.72**	0.56	0.060	159
		Distance from station (m)			
0-100	0.041**	0.78**	0.55	0.069	215
100-200	0.032	0.83**	0.54	0.064	131
200-400	0.087**	0.68**	0.44	0.060	117
400-800	0.109**	0.49**	0.40	0.064	122
800-1200	0.027	0.86**	0.67	0.058	48
1200-2000	0.161**	0.45**	0.40	0.050	42
		Soil Texture			
Clay	0.07905**	0.70793**	0.6301	0.0611	48
Clay loam	0.00792	0.81143**	0.5971	0.06211	69
Loam	0.08111**	0.58034**	0.3028	0.06012	96
Loamy sand	-0.00096	0.80678**	0.3372	0.08426	21
Sandy clay loam	-0.07746	1.28096	0.8953	0.03814	6
Sandy loam	0.04157*	0.8028**	0.5913	0.06031	129
Silt loam	0.14036**	0.55042**	0.2691	0.05989	36
Silty clay	0.08712**	0.66551**	0.4719	0.06537	270

* significant at 5% level of significance

** significant at 1% level of significance

Table A2. Linear relationship between volumetric water content (VWC) of crop fields with nearby weather stations for corn, soybean and wheat and their different growth stages.

Growth stage	Intercept (c)	Gradient (m)	r^2	RMSE	N
		Corn			
Overall	0.039	0.80**	0.46	0.069	156
V4 stage	0.227**	0.18	0.07	0.049	39
V7 stage	0.420*	-0.15	0.02	0.042	9
V10 stage	0.263	0.43	0.92	0.019	3
V11 stage	0.587*	-0.68	0.99	0.001	3
V12 stage	-0.028	0.96**	0.71	0.059	30
Tasseling	-0.021	0.97**	0.59	0.067	45
Silking	0.112	0.55	0.05	0.074	18
Grain filling	0.579	-0.95	0.78	0.014	3
Cut down	0.298	0.12	0.01	0.082	6
		Wheat			
Overall	0.076**	0.72**	0.56	0.06	159
Tillering	0.199**	0.28**	0.17	0.043	48
Jointing	0.407	-0.08	0.01	0.068	6
Flowering	0.021	0.93**	0.68	0.066	6
Hard dough	0.076**	0.70**	0.59	0.06	57
Harvested	-0.014	1.09*	0.24	0.056	18
		Soybean			
Overall	0.066**	0.69**	0.43	0.069	222
V1 stage	0.174**	0.27	0.11	0.049	30
V2 stage	0.162**	0.39*	0.15	0.075	48
V3 stage	0.183	0.39	0.81	0.004	3
V5 stage	0.327	0.16	0.02	0.045	9
V6 stage	0.08	0.65	0.09	0.094	9
Flowering	0.045	0.74**	0.51	0.077	72
Podding	0.036	0.86**	0.7	0.05	42
Pod filling	-0.003	0.87**	0.24	0.07	36

* significant at 5% level of significance
** significant at 1% level of significance

Table A3. Non-linear relationship between volumetric water content (VWC) of crop fields (N=675) with cumulative rainfall and potential evapotranspiration (PET) during the previous one to five days (D1, D2, D3, D4 and D5) for the study area.

	Intercept (c)	Gradient (m)	Gradient (m^2)	Gradient (m^3)	r^2	RMSE
Cumulative rainfall						
1 Day	0.233**	0.013**	-0.00036	2.9E-06	0.16	0.084
2 Days	0.226	0.008	-3.8E-05	-2.8E-06	0.18	0.083
3 Days	0.191**	0.159**	-0.0008**	1.4E-05**	0.27	0.079
4 Days	0.138**	0.022**	-0.001**	1.5E-05**	0.48	0.066
5 Days	0.148**	0.0156**	-0.0004**	3.9E-06**	0.49	0.066
Cumulative PET						
1 Day	0.187	0.005	0.0031	-0.00027	0.28	0.078
2 Days	0.737	-0.127	0.0111**	-0.00031**	0.29	0.078
3 Days	0.327	-0.025	0.0025	-6.8E-05*	0.28	0.078
4 Days	0.753**	-0.079*	0.0043**	-7.4E-05**	0.29	0.077
5 Days	0.836**	-0.072*	0.0031*	-4.3E-05**	0.25	0.079

* significant at 5% level of significance
** significant at 1% level of significance

Table A4. Cubical relationship between the volumetric water content (VWC) of crop fields with the four-day cumulative rainfall at different crop residue cover, crop type, distance from the station and soil texture.

	Intercept (c)	Gradient (m)	Gradient (m^2)	Gradient (m^3)	r^2	RMSE	N
Residue cover (%)							
<10	0.143**	0.0195**	-0.0007**	1.1E-05**	0.48	0.068	279
20-30	0.136**	0.0212**	-0.0009**	1.5E-05**	0.5	0.064	198
50-60	0.135**	0.0279**	-0.0013**	1.9E-05**	0.51	0.063	198
Crop type							
Alfalfa	0.966**	-0.2647*	0.0179*	-2.9E-04*	0.93	0.023	12
Corn	0.152**	0.0179**	-0.0007**	8.9E-06*	0.48	0.068	156
Dry beans	0.107**	0.0061	0.0013	-6.30E-05	0.65	0.045	33
Oats	0.023	0.0429	-0.003	7.20E-05	0.86	0.030	15
Soybean	0.137**	0.0247**	-0.0012**	1.8E-05**	0.45	0.069	222
Sugarbeet	0.061**	0.0308**	-0.0013**	1.8E-05*	0.71	0.054	45
Wheat	0.137**	0.0282**	-0.0013**	1.9E-05**	0.56	0.061	159
Distance from station (m)							
0-100	0.114**	0.0263**	-0.0012**	1.7E-05**	0.52	0.071	215
100-200	0.142**	0.0248**	-0.0011**	1.7E-05**	0.67	0.055	131
200-400	0.209**	0.0041	4.56E-05	-1.40E-06	0.37	0.064	117
400-800	0.126**	0.0293**	-0.0017**	2.9E-05**	0.44	0.062	122
800-1200	0.124**	0.0263**	-0.0009	8.30E-06	0.58	0.066	48
1200-2000	0.220**	0.0113	-0.0004	5.20E-06	0.25	0.057	42
Soil texture							
Clay	0.118**	0.0242**	-0.0007*	6.10E-06	0.75	0.057	48
Clay Loam	0.122**	0.220*	-0.0004	-4.50E-06	0.52	0.070	69
Loam	0.211**	0.0134**	-0.0004	2.80E-06	0.45	0.051	96
Loamy sand	0.230**	0.0260**	-0.0024**	6.2E-05**	0.62	0.049	21
Sandy loam	0.137**	0.0231**	-0.0009**	1.5E-05*	0.45	0.069	129
Silt loam	0.169**	0.0116	-0.0003	1.30E-06	0.45	0.050	36
Silty clay	0.188**	0.0168**	-0.0007*	9.50E-06	0.46	0.073	153
Silty clay loam	0.086**	0.0227**	-0.0007**	6.00E-06	0.68	0.048	117

* significant at 5% level of significance
** significant at 1% level of significance

Table A5. Cubical relationship between the volumetric water content (VWC) of crop fields with four-day cumulative PET at different residue content (%), crop type, distance from station and soil texture.

Residue (%)	Intercept (c)	Gradient (m)	Gradient (m^2)	Gradient (m^3)	r^2	RMSE	N
Residue cover (%)							
<10	1.281**	-0.1420*	0.0067**	-1.0E-04**	0.24	0.083	279
20-30	0.172	-0.0051	0.0012	-3.20E-05	0.37	0.071	198
50-60	0.940*	-0.1068	0.0055*	-9.2E-05**	0.31	0.076	198
Crop type							
Alfalfa	-264**	31.99**	-1.2815**	0.0170**	0.93	0.022	12
Corn	1.65**	-0.2054**	0.0099**	-1.5E-04**	0.21	0.084	156
Dry beans	-0.769	0.1237	-0.0048	5.50E-05	0.7	0.042	33
Oats	-4.96	0.6578	-0.0267	3.40E-04	0.82	0.033	15
Soybean	1.11**	-0.1328*	0.0067**	1.1E-04**	0.25	0.08	222
Sugarbeet	-4.77*	0.6064**	-0.0234**	2.97E-04**	0.68	0.056	45
Wheat	0.645	-0.059	0.0033	-5.90E-05	0.41	0.074	159
Distance from station (m)							
0-100	1.424**	-0.1741**	0.0085**	-1.3E-04**	0.33	0.084	215
100-200	0.643	-0.0709	0.0043	-7.9 E-05*	0.39	0.074	131
200-400	2.356*	-0.3004*	0.01408*	- 2.1E-04*	0.04	0.079	117
400-800	-1.709*	0.2482**	-0.0099**	1.2E-04*	0.35	0.067	122
800-1200	1.512*	-0.1854*	0.0090*	-1.4E-04**	0.57	0.066	48
1200-2000	0.79	0.1018	0.0042	5.70E-05	0.48	0.0482	42
Soil Texture							
Clay	-8.153**	1.0239**	-0.0399**	0.0005**	0.59	0.074	48
Clay Loam	-0.083	0.0379	0.0009	2.90E-07	0.59	0.065	69
Loam	-0.702	0.0978	-0.0023	1.50E-07	0.35	0.056	96
Loamy sand	-1.315	0.2414	-0.0118	1.90E-04	0.09	0.075	21
Sandy loam	4.424**	-0.6044**	0.0284**	4.30E-04	0.39	0.071	129
Silt loam	32.08*	-4.0517*	0.1705*	-0.0023	0.39	0.053	36
Silty clay	1.261**	-0.1519*	0.0076**	-1.2E-04**	0.51	0.07	153
Silty clay loam	2.051**	-0.2713**	0.0137**	-2.2E-04**	0.33	0.07	117

* significant at 5% level of significance
** significant at 1% level of significance