ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

AIDING AGRICULTURAL PRACTICES WITH THE EXPLORATION OF EARTH OBSERVATION DATA VIA MACHINE LEARNING

Ph.D. THESIS

Mehmet Furkan ÇELİK

Department of Geomatics Engineering

Geomatics Engineering Programme

AUGUST 2023

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

AIDING AGRICULTURAL PRACTICES WITH THE EXPLORATION OF EARTH OBSERVATION DATA VIA MACHINE LEARNING

Ph.D. THESIS

Mehmet Furkan ÇELİK
(501162604)

Department of Geomatics Engineering

Geomatics Engineering Programme

Thesis Advisor: Prof. Dr. Esra ERTEN

AUGUST 2023

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YER GÖZLEM UYDU VERİLERİNİN TARIMSAL UYGULAMALARA YARDIMCI OLMAK AMACIYA MAKİNE ÖĞRENME ALGORİTMALARI İLE İNCELENMESİ

DOKTORA TEZİ

Mehmet Furkan ÇELİK
(501162604)

Geomatik Mühendisliği Anabilim Dalı

Geomatik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Esra ERTEN

AĞUSTOS 2023

Mehmet Furkan ÇELİK, a Ph.D. student of ITU Graduate School student ID 501162604 successfully defended the thesis entitled "AIDING AGRICULTURAL PRACTICES WITH THE EXPLORATION OF EARTH OBSERVATION DATA VIA MACHINE LEARNING", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Esra ERTEN** Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Orhan AKYILMAZ** Istanbul Technical University	.....
	Prof. Dr. Erchan APTOULA Sabanci University	.....
	Prof. Dr. Nebiye MUSAOĞLU Istanbul Technical University	.....
	Prof. Dr. Koray KAYABOL Gebze Technical University	.....

Date of Submission : **22 June 2023**
Date of Defense : **4 August 2023**

To my lovely family,

FOREWORD

No success can be the result of individual work. That's why my greatest gratitude goes to my mother, father, and sisters. I would not have been able to complete this research without feeling in my heart the faith of my family, who have encouraged and supported me throughout my life.

I am sincerely grateful to conduct this research under the supervision of Prof. Dr. Esra ERTEN, who always supports me with her expertise. Her experience and guidance broadened my academic perspective from the beginning. I consider myself extremely lucky to have had the chance to work under her supervision. I would like to thank Prof. Dr. Orhan AKYILMAZ and Prof. Dr. Erchan APTOULA, members of the thesis committee, for their feedback and insightful guidance throughout the development of this thesis. I would like to express my thanks to Prof. Dr. Gustau CAMPS-VALLS for accepting my collaboration and for his invaluable contributions to my academic perspective. His expertise and constructive feedback have been critical to the success of this thesis.

I would like to express my special gratitude to Mustafa Serkan IŞIK and Ozan ÖZTÜRK, with whom I have had the pleasure of working as colleagues and am honored to call friends. Their friendship, guidance, and encouragement significantly contributed to the successful completion of my thesis. I would also like to thank Elif DEMİR ÖZBEK, Serpil ATEŞ AYDAR, Umut AYDAR and Hüseyin MERCAN who have been both friends and mentors to me since my bachelor's degree.

This thesis is supported by the Scientific Research Projects Coordination of Istanbul Technical University under Project Number MDK-42745. The support received through Application Number: 1059B141900633 from the Scientific and Technological Council of Turkey (TÜBİTAK) via the International Research Fellowship Programme for Ph.D. Students (2214-A) is gratefully acknowledged.

As a final word, I would like to extend my thanks to each and every person who supported me with their friendship during this period.

August 2023

Mehmet Furkan ÇELİK
(Geomatics Engineer, M.Sc.)

TABLE OF CONTENTS

ABBREVIATIONS

A : Aspect
ANN : Artificial Neural Network
API : Application Programming Interface
BBCH : Biologische Bundesanstalt, Bundessortenamt and Chemical Industry
BD : Bulk Density
CEC : Cation Exchange Capacity
CONUS : Continental United States
DL : Deep Learning
DLD : Daylight Duration
DT : Decision Tree
EBM : Explainable Boosting Machine
ESA : European Space Agency
ET : Evapotranspiration
EVI : Enhanced Vegetation Index
FPAR : Fraction of Photosynthetically Active Radiation
GEE : Google Earth Engine
GAMs : Generalized Additive Models
GA^{2}Ms : Generalized Additive Models plus Interactions
GSA : Global Sensitivity Analysis
H : Elevation
HS : Hillshade
ISMN : International Soil Moisture Network
LAI : Leaf Area Index
LASSO : Least Absolute Shrinkage and Selection Operator
LightGBM : Light Gradient-Boosting Machine
LOO : Leave-One-Out
LST_D : Daytime Land Surface Temperature
LST_N : Nighttime Land Surface Temperature
LSTM : Long Short Term Memory
MAE : Mean Absolute Error
MAPE : Mean Absolute Percentage Error
ML : Machine Learning
MODIS : Moderate Resolution Imaging Spectroradiometer
MSE : Mean Square Error
N : Total Nitrogen
NCA : Neighbourhood Component Analysis
NDVI : Normalized Difference Vegetation Index
P : Precipitation
PCE : Polynomial Chaos Expansion
PolSAR : Polarimetric Synthetic Aperture Radar

R^2 : Coefficient of Determination
RLR : Ridge Linear Regression
RMSE : Root Mean Square Error
RNN : Recurrent Neural Network
S : Slope
S1 : Sentinel-1
S2 : Sentinel-2
SAR : Synthetic Aperture Radar
SGD : Stochastic Gradient Descent
SM : Soil Moisture
SMAP : Soil Moisture Active Passive
SMOS : Soil Moisture and Ocean Salinity
SR : Solar Radiation
T : Temperature
T_{max} : Maximum Air Temperature
T_{min} : Minimum Air Temperature
ubRMSE : Unbiased Root Mean Square Error
USDA : United States Department of Agriculture
XGBoost : Extreme Gradient Boosting

SYMBOLS

$\mathscr{M}(\boldsymbol{X})$	: The function defined by multivariate orthogonal polynomial
$\Psi_{\boldsymbol{\alpha}}$	: The multivariate orthogonal polynomial basis
$y_{\boldsymbol{\alpha}}$	: The coefficient of the polynomials
N	: The dimension of the input feature vector
M	: The truncation level
$\phi_{\alpha_i}^{(i)}$	: The univariate orthogonal polynomial with the truncate set of multi indices α_i
d	: The total polynomial order
argmin	: Minimization function
ε_{LOO}	: Leave-one-out error
$\hat{\mu}_Y$	: The sample mean of the validation set
S_i	: First order Sobol Indices
$S_{i,j}$	: Second order Sobol Indices
$S_{1,2,\ldots,n}$	: High order Sobol Indices
Var[...]	: Variance
$\lvert HH\rvert^2$	: Horizontal-Horizontal backscatter coefficient
$\lvert HV\rvert^2$	: Horizontal-Vertical backscatter coefficient
$\lvert VV\rvert^2$	: Vertical-Vertical backscatter coefficient
H	: Entropy
A	: Anisotropy
α	: Alpha
$\phi_{HH,VV}$	: Phase differences between HH and VV
$\phi_{HH,HV}$	: Phase differences HH and HV
$\phi_{VV,HV}$	: Phase differences VV and HV
$\rho_{HH,VV}$	: Correlation between HH and VV
$\rho_{HH,HV}$	: Correlation between HH and HV
$\rho_{VV,HV}$	: Correlation between VV and HV
$\lvert HH+VV\rvert^2$	: 1^st Pauli Component
$\lvert HH-VV\rvert^2$	: 2^nd Pauli Component
$\phi_{HH+VV,HH-VV}$	: 1^st & 2^nd Pauli components phase differences
$\rho_{HH+VV,HH-VV}$	: 1^st & 2^nd Pauli components correlation
i_t	: The input gate of the LSTM cell
f_t	: The forget gate of the LSTM cell
o_t	: The output gate of the LSTM cell
c_t	: The cell state for the LSTM cell
w_i	: The weight matrix of the input gate
w_f	: The weight matrix of the forget gate
w_o	: The weight matrix of the output gate
w_c	: The weight matrix of the cell state
h_{t-1}	: The hidden state from previous cell

b_i : The bias vector of input gate
b_f : The bias vector of forget gate
b_o : The bias vector of output gate
b_c : The bias vector of cell state
σ : The sigmoid activation function
$\odot$: Element wise multiplication
$tanh$: The hyperbolic tangent activation function
y_i : The actual value
$\hat{y}_i$: The predicted value
$\bar{y}_i$: The mean value of the actual value
w : The sequential days (window size)
β_0 : The intercept
f_i : The each feature function
n : Feature dimension
x_i : The input features
g : The link function
$f_{i,j}$: The pairwise interactions of input features
E[...] : Expectation operator

LIST OF TABLES

LIST OF FIGURES

AIDING AGRICULTURAL PRACTICES WITH THE EXPLORATION OF EARTH OBSERVATION DATA VIA MACHINE LEARNING

SUMMARY

The rapid growth of the global population, coupled with the decline in available agricultural fields, the effects of climate change, and soil degradation, pose significant threats to food security. As the population continues to rise, the demand for food and agricultural products increases, putting pressure on optimizing limited resources and their use. The human-made global climate crisis, primarily driven by fossil fuel emissions, worsens the issue, causing extreme weather events and displacing communities. Minimizing environmental damage and maximizing agricultural efficiency is crucial for ensuring a sustainable supply of essential food resources and the well-being of humanity. Obtaining accurate information about agriculture is vital for decision-makers, but traditional in-situ measurements are insufficient to represent the fields and are time-consuming. Remote sensing satellite images provide a solution by offering comprehensive and reliable data, overcoming the limitations of traditional methods, and enabling effective monitoring of agricultural fields on a regional or larger-scale level. Remote sensing satellite imaging technologies, including synthetic aperture radar (SAR) and multi-spectral imaging (MSI) satellites, provide valuable information for Earth Observation (EO) studies. SAR satellites are able to operate in any weather condition, day or night, and penetrate cloud cover, making them highly effective for monitoring Earth's surface. Despite their reliance on clear skies and solar energy, MSI satellites play a crucial role in agricultural monitoring due to their value with a wide range of spectral bands. Both satellite systems have a significant role in observing agricultural fields; SAR satellites are sensitive to detecting morphological changes in crops and MSI satellites have the capability to monitor chemical changes in vegetation. The satellite images offer insights into crop health, growth stages, and potential yield prediction through parameters derived from MSI and SAR images. Utilizing machine learning (ML) algorithms to analyze remote sensing data for agricultural research has opened up a wide range of possibilities for conducting comprehensive studies based on the ability of these algorithms to grasp nonlinear relationships associated with electromagnetic radiation and vegetation. Agricultural planning authorities and researchers can obtain critical insights into many aspects of agriculture and make informed decisions by utilizing the power of these advanced computing approaches. For this purpose, in order to address the critical challenges in monitoring agricultural fields and understanding the interrelation between environmental factors and agricultural activities, three-stage research that implements state-of-art ML and deep learning (DL) methods on remote sensing images has been conducted within the scope of this thesis. These challenges include various aspects of agricultural analysis and can be effectively tackled using the power of ML and DL algorithms that explain the models' behavior in an easy format to understand.

In the first study, regression analysis was used to examine the estimation of biophysical parameters using only SAR remote sensing satellite data. Among the regression methods, polynomial chaos expansion (PCE) is one of the reliable and interesting ones due to its tight relationship with uncertainty quantification. One of the advantages of PCE is that global sensitivity analysis (GSA) with Sobol's method can be analytically computed from polynomial coefficients if the input space is statistically independent. However, most of the phenomena include dependent features, either statistically or physically. Therefore, an independent and uncorrelated input space must be created before the regression analysis. In this paper, we performed PCE-based regression analysis for the estimation of biophysical parameters of crops. The study was conducted in the experimental fields of field pea, barley, canola, and oat of the AgriSAR2009 campaign. The input parameters of the regression model were formed by creating polarimetric features derived from RADARSAT-2 imagery. The estimated biophysical parameters were based on the discrete in-situ measurements of leaf area index (LAI) and normalized difference vegetation index (NDVI), scattered semi-randomly in each crop field. We implemented neighborhood component analysis (NCA) to create an independent and uncorrelated input space by eliminating correlations. Once the model was created, we investigated the importance of features that drive the PCE-based regression models applying GSA with Sobol's method. Besides the individual effects of each feature, their interactions were found to be significant.

In the second study, time series analysis was conducted to obtain short-term soil moisture in field scale, integrating satellite imaging, climate, and auxiliary data. The recent advancements in different types of satellite imagery coupled with deep learning-based frameworks have paved the way for large-scale SM estimation. This research combined high spatial resolution Sentinel-1 (S1) backscatter data and high temporal resolution Soil Moisture Active Passive (SMAP) SM data to create short-term SM predictions that can accommodate agricultural activities. We created a deep learning model to forecast the daily SM values using time series of climate and radar satellite data, soil type, and topographic data. The model was trained with static and dynamic features that influence SM retrieval. While the topography and soil texture data were taken as stationary, SMAP SM data and S1 backscatter coefficients, including their ratios and climate data were fed to the model as dynamic features. As a target data to train the model, we used *in-situ* measurements acquired from the International Soil Moisture Network (ISMN). We employed a deep learning framework based on Long Short-Term Memory (LSTM) architecture with two hidden layers with 32 unit sizes and a fully connected layer. The model's performance was also evaluated concerning above-ground biomass, land cover classes, soil texture variations, and climate classes. The model prediction ability was lower in areas with high normalized difference vegetation index (NDVI) values. Moreover, the model can predict better in dry climate areas, such as arid and semi-arid climates, where precipitation is relatively low. The daily prediction of SM values based on microwave remote sensing data and geophysical features was successfully achieved using an LSTM framework to assist various studies such as hydrology and agriculture.

In the third study, the importance of the input features was investigated during the cotton phenological cycle in order to predict yield using an explainable artificial

intelligence. The potential cotton yield can be predicted by integrating the climatic factors, soil parameters, and biophysical parameters observed by high temporal and spatial resolution remote sensing satellites. This study used a multisource dataset to create an explainable and accurate predictive model for cotton yield prediction over the continental US (CONUS). A recently proposed glass-box method called Explainable Boosting Machine (EBM), which provides transparency, reliability, and ease of interpretation, was implemented. Accuracy performance was compared with well-known ML methods for predicting cotton yields. The EBM showed higher accuracy against other glass-box methods and competitive results with black-box models. With the help of the EBM, the importance of individual features and their pairwise interactions was revealed without applying any post-hoc methods. The study findings showed that the precipitation (P), enhanced vegetation index (EVI), and leaf area index (LAI) are the three most important dynamic features. The dynamic features are the driver of the created model with 78% of the overall feature importance, followed by pairwise interactions of the features with 16% contribution. Lastly, static features contribute 6% to the overall feature importance. The study highlights the importance of using multisource data and interactions of the input features and providing an interpretable model to understand the inner dynamics of cotton yield predictions.

YER GÖZLEM UYDU VERİLERİNİN TARIMSAL UYGULAMALARA YARDIMCI OLMAK AMACIYA MAKİNE ÖĞRENME ALGORİTMALARI İLE İNCELENMESİ

ÖZET

Küresel nüfusun hızla artması, mevcut tarım alanlarının azalması, iklim değişikliği ve toprak bozulmasının etkileri birlikte değerlendirildiğinde gıda güvenliğinin tehdit altında olduğu ortaya çıkmaktadır. Nüfus artmaya devam ettikçe gıda ve tarım ürünlerine olan talep artmakta, bu da sınırlı kaynakların optimize edilmesini ve verimli kullanımını zorunlu kılmaktadır. Başta fosil yakıt emisyonlarının neden olduğu insan kaynaklı küresel iklim krizinin, olağanüstü hava olaylarına neden olması ve büyük insan topluluklarını yerinden göç etmeye zorlaması sorunu daha da kötüleştirmektedir. Çevreye verilen zararın en aza indirgenmesi ve tarımsal verimliliğin en üst düzeye çıkarılması, temel gıda kaynaklarının sürdürülebilir bir şekilde tedarik edilmesi insanlığın refahı için hayati önem taşımaktadır. Tarımsal bitkilerin gelişim süreci ve üretim verimliliği hakkında doğru bilgi edinmek karar alıcılar için oldukça önemlidir. Ancak geleneksel olarak arazide ve sensör tabanlı yapılan ölçmeler tarlaların ve tarımsal bitkilerin gözlenmesi için gerekli olan parametrelerin modellenmesinde yetersiz kalmakta ve zaman almaktadır. Uzaktan algılama uydu görüntüleri, kapsamlı ve güvenilir veriler sunarak, geleneksel yöntemlerin sınırlamalarını aşmakta ve tarım alanlarının bölgesel veya daha büyük ölçekte izlenmesini sağlayarak etkili bir çözüm sunmaktadır. Yapay açıklıklı radar (SAR) ve çok bantlı görüntüleme (MSI) uyduları dahil olmak üzere uzaktan algılama uydu görüntüleme teknolojileri, Yer Gözlem (EO) çalışmaları için değerli bilgiler sağlar. SAR uyduları gece veya gündüz her türlü hava koşulunda çalışabilmekte ve bulut örtüsünü aşarak Dünya yüzeyinin izlenmesinde son derece etkili olmaktadır. Açık ve bulutsuz gökyüzüne ve güneş enerjisine bağımlı olmalarına rağmen, MSI uyduları çeşitli spektral bantlara sahip olmaları nedeniyle tarımsal izlemede önemli bir rol oynamaktadır. Her iki uydu sistemi de tarım alanlarının gözlemlenmesinde etkili çözümler sunabilmektedir. SAR uyduları tarımsal bitkilerdeki morfolojik değişiklikleri tespit etme konusunda başarılı iken, MSI uyduları bitki örtüsündeki kimyasal değişiklikleri izleme kapasitesine sahiptir. MSI ve SAR uzaktan algılama uydu görüntülerinden elde edilen biyofiziksel parametreler, bitki sağlığı, büyüme evreleri ve potansiyel verim tahmini hakkında bilgi edinmemize olanak sağlayabilmektedir. Tarımsal araştırmalar için uzaktan algılama verilerini analiz etmek üzere makine öğrenimi (ML) algoritmaları kullanılmaktadır. ML algoritmaları, elektromanyetik radyasyon ve bitki örtüsü ile doğrusal olmayan ilişkileri kavrama yeteneğine dayanan kapsamlı çalışmalar yürütmek için geniş bir araştırma alanı yaratmıştır. Gelişmiş hesaplama yaklaşımlarının gücünü kullanarak tarımsal üretim hakkında kritik bilgiler, tarımsal planlama yetkilileri ve araştırmacılar tarafından elde edilebilmekte ve bilinçli karar almalarına olanak sağlayabilmektedir. Bu amaçla, tarım alanlarının izlenmesi ve çevresel faktörler ile tarımsal faaliyetler arasındaki karşılıklı ilişkinin anlaşılmasındaki zorlukları ele almak için, bu tez

kapsamında uzaktan algılama görüntüleri üzerinde ML ve derin öğrenme (DL) yöntemlerini uygulayan üç aşamalı bir araştırma yürütülmüştür. Tarımsal üretim için gerekli olan farklı parametreler geliştirilen modellerin iç dinamiklerini açıklayan ML ve DL yöntemlerini kullanılarak incelenmiş ve değerlendirilmiştir.

İlk çalışmada SAR uzaktan algılama uydu görüntülerinden yararlanılarak bitki biyofiziksel parametrelerinin regresyon analizi ile tahmini gerçekleştirilmiştir. Polinomsal kaos açılımı (PCE) belirsizlik ölçümü gerçekleştirilebilmesi nedeniyle regresyon yöntemleri arasında dikkat çekmektedir. PCE yöntemi ile kurulan fonksiyonel modelde, girdi vektörlerinin birbirleri ile korelasyonsuz olması durumunda, denklem katsayıları kullanılarak küresel duyarlılık analizi (GSA) yapılabilmekte ve Sobol İndisleri hesaplanabilmektedir. Fakat yer bilimleri açısından incelenen fenomenleri etkileyen girdi vektörleri istatiksel ya da fiziksel bağımlı veriler içermektedir. Bu nedenle PCE regresyon analizininden önce birbiri ile korelasyonsuz bir girdi vektör kümesi oluşturulmalıdır. Bu çalışmada tarımsal ürünlerin biyofiziksel parametrelerinin tahmini için PCE tabanlı regresyon analizi gerçekleştirilmiştir. Kullanılan veri seti AgriSAR2009 kampanyası dahilinde biyofiziksel parametleri toplanan bezelye, arpa, kanola ve yulaf tarımsal bitkileri içermektedir. Tahmini yapılacak olan biyofiziksel parametreler, her bir tarlanın dört bölgesinde yapılan yaprak alanı indeksi (LAI) ve normalleştirilmiş fark bitki örtüsü indeksi (NDVI) ölçümlerine dayanmaktadır. Regresyon modelinin girdi parametreleri ise RADARSAT-2 görüntülerinden türetilen polarimetrik özellikler ile oluşturulmuştur. Girdi parametrelerini oluşturan polirimetrik özellikler arasından korelasyonlu olanları eleyebilmek ve alt küme girdi seti oluşturabilmek için komşuluk bileşenleri analizi (NCA) uygulanmıştır. PCE regresyon modeli ile fonksiyonel ilişki kurulması ile Sobol yöntemiyle GSA yapılarak modeli yönlendiren girdi vektörlerinin önemini araştırılmıştır. Her bir girdi vektörünün tekil etkisinin yanı sıra birbiri ile etkileşiminin de regresyon analizinde önemli olduğu sonucuna varılmıştır.

İkinci çalışmada uzaktan algılama uydu verisi, iklim parametreleri ve yardımcı veriler kullanılarak arazi ölçeğinde kısa vadeli toprak nemi tahmini için zaman serisi analizi gerçekleştirilmiştir. Uzaktan algılama uydu teknolojilerindeki gelişmeler ve derin öğrenmeye dayalı analizler geniş ölçekli toprak nemi tahmini çalışmalarına olanak sağlamaktadır. Bu çalışmada, tarımsal faaliyetlere yardımcı veri sağlamak için yüksek mekansal çözünürlüklü Sentinel-1 (S1) geri saçılım verileri ve yüksek zamansal çözünürlüklü Soil Moisture Active Passive (SMAP) toprak nemi verilerinin birlikte değerlendirilmesi ile kısa vadeli toprak nemi tahmini gerçekleştirilmiştir. Çalışma kapsamında toprak nemini etkileyen sabit parametreler ile zamanla değişen parametreler birleştirilerek model içerisinde birlikte değerlendirilmiştir. Günlük toprak nemi değerlerinin tahmin edilebilmesi için, topografya ve toprak dokusu verileri sabit olarak alınırken, SMAP toprak nemi verileri, S1 geri saçılım katsayıları ve oranları ile iklim verileri dinamik özellikler olarak kullanılarak derin öğrenme modeli oluşturulmuştur. Modelin eğitiminde çıktı verisi olarak, Uluslararası Toprak Nemi Ağı'na (ISMN) ait toprak nemi ölçme sensörlerinden elde edilen veriler kullanılmıştır. Derin öğrenme modeli olarak, iki katmanlı ve 32 nörona sahip Long Short-Term Memory (LSTM) mimarisine ek olarak tek nörona sahip doğrusal katman eklenmiştir. Eğitilen modelin doğruluk performansı toprak üstü biyokütle, arazi örtüsü sınıfları, toprak dokusu ve iklim sınıfları açısından ayrı ayrı değerlendirilmiştir. Modelin

tahmin kabiliyetinin normalleştirilmiş fark bitki örtüsü indeksi (NDVI) değerlerinin yüksek olduğu alanlarda daha düşük olduğu gözlemlenmiştir. Ayrıca model, yağışın nispeten düşük olduğu kurak ve yarı kurak iklimler gibi kuru iklim bölgelerinde daha iyi tahmin yapabilmektedir. Mikrodalga uzaktan algılama verilerine ve jeofiziksel özelliklere dayalı, toprak nemi değerinin günlük tahmini, hidroloji ve tarım gibi çeşitli çalışmalara yardımcı olmak için LSTM derin öğrenme modeli kullanılarak başarıyla gerçekleştirilmiştir.

Üçüncü çalışmada, açıklanabilir yapay zeka algoritması ile pamuk fenolojik döngüsü sırasında tarımsal ürün rekoltesi tahmin etmek için kullanılacak girdi parametrelerinin önemi araştırılmıştır. Yüksek zamansal ve mekansal çözünürlüklü uzaktan algılama uyduları kullanılarak gözlemlenen biyofiziksel parametreler, iklimsel faktörler ve toprak parametrelerinin entegrasyonuyla potansiyel pamuk verimi tahmin edilmiştir. Bu çalışmada, Kıtasal Amerika Birleşik Devletleri'nde (CONUS) pamuk verimi tahmini amacıyla açıklanabilir ve doğru bir tahmin modeli oluşturmak için çok kaynaklı bir veri kümesi kullanılmıştır. Şeffaflık, güvenirlik ve yorumlama kolaylığı sağlayan Explainable Boosting Machine (EBM) cam-kutu yöntemi uygulanmıştır. Pamuk verim tahmininin doğruluk performansının kıyaslanabilmesi için yaygınlıkla kullanılan ML algoritmaları test edilmiştir. EBM, diğer cam-kutu yöntemlere göre daha yüksek doğruluk sağlamış ve kara-kutu modellerle kıyaslanabilir sonuçlar göstermştir. EBM'nin yardımıyla her bir özelliğin ve ikili etkileşimlerinin önemi ek bir hesap yükü gerektirmeden ortaya konabilmiştir. Çalışma bulgularına göre yağış (P), artırılmış bitki örtü indeksi (EVI) ve yaprak alan indeksi (LAI) en önemli üç dinamik özellik olarak ortaya çıkmıştır. Dinamik girdi vektörleri, toplam girdi vektörleri öneminin %78'ini oluşturarak modelin itici gücü olmuş ve bunu %16'lık bir katkıyla girdi vektörleri arasındaki ikili etkileşimler takip etmiştir. Son olarak, statik girdi vektörlerinin toplam girdi vektörleri önemine %6'lık bir katkıda bulunduğu ortaya çıkarılmıştır. Çalışma, çok kaynaklı veri ve girdi özelliklerinin etkileşimlerinin kullanımının ve pamuk verim tahmini yapılırken kullanılan ML yönteminin iç dinamiklerini anlamak için yorumlanabilir bir modelin önemini vurgulamaktadır.

1. INTRODUCTION

Nowadays, agricultural production has emerged as a pressing global issue that requires immediate attention, and it is expected to cause a global food crisis in the upcoming years. The drastic growth of the world population, coupled with diminishing agricultural land, the effects of global climate change, and the degradation of soil, pose substantial threats to human beings' access to essential agricultural products. During the 1970s, the global population was approximately 3.7 billion, with agricultural production utilizing an area of 45.5 million square kilometers. By the year 2000, the world population had risen to 6.2 billion, and agricultural land had expanded to cover 48.7 million square kilometers. However, as we approach 2020, the global population has reached 8 billion, yet the agricultural areas in use have decreased to 47.4 million square kilometers [1,2]. Furthermore, based on current projections, it is expected that there will continue to be an inverse relationship between population growth and the decrease in available agricultural land in the future. As the global population steadily increases, the demand for food and agricultural products also rises. This places additional pressure on the limited resources of agricultural land. In addition to these facts, global climate change and the degradation of soil due to human activities and consumption desires have magnified the issue, making it even more challenging to address. According to the United Nations, fossil fuels which are mainly used as energy sources for industry, transportation, and powering buildings, cause 75% of global greenhouse emissions, which makes it the main driver for the global climate crisis. This crisis, which is completely man-made, causes extreme weather events such as extremely high air temperatures, floods, and drought, as well as displacement of humans. While all these events are interconnected in a cause-and-effect relationship, it becomes crucial to minimize damage to the Earth while maximizing the efficiency of agricultural lands. This approach is necessary to ensure the provision of essential food resources at a certain standard, sustaining the well-being and continuity of humanity.

Accurate and reliable information obtained from the agricultural domain is crucial in aiding decision-makers to program effective agricultural practices. However, acquiring such information through in-situ measurements in agricultural fields presents substantial challenges in terms of complexity, time consumption, and financial burden. Traditional methods fail to deliver sufficient information for regional or large-scale studies, even when dealing with comparatively small areas of arable land. Fortunately, remote sensing satellite images offer a solution by providing comprehensive temporal and spatial data integrity when observing agricultural fields to overcome this obstacle.

Earth observation studies can benefit from the valuable information provided by remote sensing satellite imaging technologies, which operate in various electromagnetic spectrum regions. Synthetic aperture radar (SAR) satellites, one of these technologies, employ microwave signals, while multi-spectral imaging (MSI) satellites operate in the visible and infrared ranges. Compared to one another, SAR and MSI satellite imagery systems each offers benefits and drawbacks in terms of investigating the Earth's surface and its many features. SAR imaging satellites distinguish by their ability to operate day and night in almost any weather conditions and to penetrate cloud cover, making them highly effective for earth surface monitoring. SAR systems acquire information in a particular wavelength through backscatter intensity and polarization, whereas MSI captures data across multiple spectral bands. On the other hand, MSI satellites heavily rely on clear skies and solar power to monitor the Earth's surface. Nevertheless, both satellites play a valuable role in observing agricultural fields.

SAR satellites are sensitive to the dielectric constant and morphological changes in crops, while MSI satellites allow for the monitoring of chemical changes. These satellite images provide valuable insights into crops' health and growth stages at any time and enable the calculation of potential yield. The utilization of parameters derived from MSI images, including Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (FPAR), Normalized Difference Vegetation Index (NDVI), and Enhanced Vegetation Index (EVI), as well as polarimetric features obtained from SAR images, enables the analysis of the physical and chemical transformations occurring throughout the phenological stages of agricultural products. Moreover, soil moisture (SM) is another critical biophysical parameter significantly influencing crop

development. It plays a vital role in regulating soil and air temperature, directly impacting photosynthesis and energy production in crops. Additionally, SM is closely linked to root development, facilitating access to essential nutrients in the soil. These factors collectively contribute to the overall growth and development of crops. SM is influenced by several factors, including topography, soil texture, and climate data. These factors interact to determine the amount of moisture present in the soil. The specific effects of climate data, such as precipitation, temperature, and evapotranspiration, vary depending on the topography and soil texture. For instance, even in regions with the same elevation and soil texture, SM can vary due to different slopes. Similarly, areas with similar topographic characteristics but different soil textures will have varying SM levels due to the soil's water-holding capacity differences.

From the point of view of agricultural field monitoring, there is a complex relationship between a multidimensional input space and output. While the biophysical parameters can be obtained from satellite images, the complex relationship between the SM, climate data, and soil properties that impact the crops' growth and, eventually, the potential yield to be produced needs to be solved. Exploring the interrelation between multidimensional input features and their impact on agricultural production and highlighting the significance of complex relationships in managing and planning agricultural practices.

In the present era, the advancement of satellite technologies has provided us with an abundance of data to facilitate targeted observations. This thesis addresses the challenge of effectively monitoring agricultural fields, obtaining biophysical parameters, and determining crop yields by leveraging data-driven methods and integrating existing data sources. The functional relationship between multidimensional input space and output, which is non-linear in nature in the case of agricultural monitoring, has been analyzed using various artificial intelligence (AI) models. However, the focus of research today is not only on achieving high accuracy with AI methods but also on explaining this functional relationship and determining drivers of the established model. It is difficult for decision-makers to plan without understanding the internal dynamics of the model. Therefore, whatever artificial

intelligence model is used, it should be explainable and interpretable. While some AI methods, such as glass-box methods, can be explained by the coefficients of the functional models, black-box methods require post-hoc methods to be explained and interpret the effective input features.

In the first chapter of this thesis, two crucial biophysical parameters, LAI and NDVI, were estimated from full polarimetric SAR images. LAI serves as a physical parameter that reflects the extent of crop growth by measuring the coverage of the crop canopy on the unit surface area. On the other hand, NDVI functions as a chemical parameter that offers insights into the crop's health by indicating the chlorophyll content present within the crop. These parameters are commonly generated from MSI satellite images. However, the MSI systems are sensitive to atmospheric conditions and cloud coverage, which can limit obtaining data, particularly in areas with frequent cloud cover, precipitation, or haze. As a result, there is a lack of comprehensive information regarding the biophysical properties of crops in such regions. Therefore, the biophysical parameters produced from MSI imagery, like LAI and NDVI, are intended to be estimated from SAR satellite imagery, which can collect data independently of atmospheric conditions, cloud cover, and solar power. For this purpose, the data set obtained during the AgriSAR2009 campaign funded by ESA was used in this study. The data set collected in the Indian Head region was evaluated in this research, which covers eight fields with four crop types: field pea, barley, canola, and oat. In a single growth period covering June to July, fully polarimetric RADARSAT-2 images were acquired, and the in-situ measurements, including LAI and NDVI, were collected in the four corners of each field and were conducted during the corresponding growing cycle. In order to determine the output value corresponding to each pixel, the inverse distance weighting interpolation method was applied to the NDVI and LAI parameter. A regression model was then built to identify NDVI and LAI parameters within the test samples using a total of 23 polarimetric features created from the fully polarimetric images. As a regression method, the polynomial chaos method, which allows the functional relationship between input and output to be established by orthogonal polynomials, is preferred. The PCE method includes the interaction between higher-order polynomials of the input parameters, which improves

the regression performance, and coefficients of the polynomials can be directly used to conduct sensitivity analysis.

In the second chapter of this thesis, time series analysis incorporating multisource data integration techniques was employed to estimate SM, a crucial biophysical parameter that significantly influences crop growth and health. SM helps crop root development and allows nutrients and water in the soil to be transferred to the shoot system. This allows to produce energy through photosynthesis and enables the crop to be productive. SM, an essential parameter for crop growth and health, can be monitored by SAR satellite missions that are sensitive to the amount of water in the soil, which affects the dielectric constant. The two common satellite missions dedicated to measuring SM content are SMAP and SMOS, which cover quasi-globally and acquire data with high temporal and low spatial resolutions. The SAR satellite mission, Sentinel-1 (S1), can also be utilized to estimate SM. However, it is important to note that neither the high spatial resolution of the Sentinel-1 mission nor the high temporal resolution of the SMAP and SMOS missions can provide daily SM data suitable for agricultural activities at the field scale, especially concerning irrigation. Consequently, accurate SM prediction becomes crucial in addressing this limitation and supporting effective agricultural practices. Due to its successful performance with sequential data, time series analysis for daily SM prediction was performed using the LSTM method to provide decision-makers with daily and field-scale SM. This research aims to improve short-term SM prediction at the field scale by combining high temporal resolution SMAP SM product, high spatial resolution S1 backscatter coefficients, and auxiliary data relevant to agricultural activities. We incorporated SM data from ground stations of the International Soil Moisture Network (ISMN) located worldwide to accomplish this. An LSTM model was trained using two microwave radar datasets (SMAP and S1), along with soil texture, climate, and topographical data, which served as predictors for SM. We successfully achieved short-term SM forecasts at a field scale by leveraging microwave remote sensing satellite-based observations. The model employed in this study accurately predicted SM values for the following day, offering high spatial resolution across regions characterized by various geophysical properties and climate classes.

The third chapter of this thesis focuses on yield estimation based on biophysical parameters, climate data, and soil properties for cotton, which is vital in supporting the textile industry and has great power in the global economy. Making efficient agricultural decisions based on understanding the spatial and temporal variations of cotton yield and their relation to changes in climatic and pedological conditions remains challenging. Developing a reliable yield estimation model that can assist farmers in agricultural planning requires an explicit interpretation of the functional relationship between biophysical parameters, climate data, soil properties, and cotton yield. To this end, the Continental United States (CONUS) was chosen as the study area because the United States Department of Agriculture (USDA) thoroughly investigates cotton production in the CONUS and releases a high-resolution crop classification map with yield records annually. A comprehensive set of 18 features was collected to analyze and comprehend the relationship between yield records and their predictors. These features included readily available MODIS products such as EVI, LAI, and FPAR, as well as land surface temperature during daytime (LST_D) and nighttime (LST_N). Additionally, surface SM data from SMAP and climate data from Daymet V4, including precipitation, maximum and minimum air temperature, solar radiation, and daylight duration, were incorporated. Furthermore, seven soil properties obtained from SoilGrid, including sand, silt, clay content, bulk density, cation exchange capacity, nitrogen, and pH, were considered predictors. The integration of these various features aimed to establish a comprehensive understanding of the functional relationship between yield records and their associated factors. The explainable boosting machine (EBM) method, which also considers the pairwise interaction between input vectors, is preferred for establishing the functional relationship between input and output vectors. The preference for EBM stems from their ability to provide interpretable and explainable results. These models are transparent in nature, eliminating the need for any post-hoc methods to understand the underlying model. The EBM demonstrated model accuracies similar to other boosted tree algorithms based on the yield prediction results obtained. The analysis identified the importance of biophysical and climatic parameters at specific phenological stages and soil properties that influenced cotton yield within the model. Moreover, a

monthly-based implementation of EBM has been conducted, revealing that cotton yield estimation can be achieved with high accuracy in advance within the season.

Following the introduction, this thesis is structured around three articles, as mentioned above, and chapters 2, 3, and 4 are devoted to each relevant article. The final section concludes the thesis by presenting the discussions and conclusions from the articles. Three chapters of this thesis have been published in the following SCI-Expanded indexed journals;

Chapter 2 is published as a paper entitled "*Biophysical Parameter Estimation of Crops from Polarimetric Synthetic Aperture Radar Imagery with Data-Driven Polynomial Chaos Expansion and Global Sensitivity Analysis*" in Computers and Electronics in Agriculture.

Chapter 3 is published as a paper entitled "*Soil Moisture Prediction from Remote Sensing Images Coupled with Climate, Soil Texture and Topography via Deep Learning*" in Remote Sensing.

Chapter 4 is published as a paper entitled "*Explainable Artificial Intelligence for Cotton Yield Prediction with Multisource Data*" in IEEE Geoscience and Remote Sensing Letters.

2. BIOPHYSICAL PARAMETER ESTIMATION OF CROPS FROM POLARIMETRIC SYNTHETIC APERTURE RADAR IMAGERY WITH DATA-DRIVEN POLYNOMIAL CHAOS EXPANSION AND GLOBAL SENSITIVITY ANALYSIS[1]

2.1 Introduction

The increasing world population, urbanization, and natural disasters caused by climate change encourage us to continuously observe, manage, and plan agricultural production. Therefore, precision agriculture practices are increasing day by day and maintain as a hot topic. However, with traditional methods, it is still inefficient in terms of both cost and time to measure the crop's biophysical parameters, phenological stages, productivity, and health. Moreover, it is almost impossible to obtain biophysical parameters for the entire agricultural field, especially in large-scale industrial agriculture.

In this context, monitoring agricultural fields and estimating biophysical parameters of crops from satellite data allow us to plan and manage precision agriculture practices for large-scale agricultural applications. Satellite data are essential components in agricultural studies and play a key role in precision agriculture. Sensitivity to the physical properties of the target object, day & night imaging capability, and less impact of weather conditions on radar signals make synthetic aperture radar (SAR) satellite sensors a suitable data source for agricultural field monitoring [3]–[5].

SAR satellite data are frequently used for classification and for regression of biophysical parameters of agricultural products. Previous studies have explored the relationships between phenological stages and polarimetric features derived from SAR data. In the literature, while some of the studies present the relationship

[1]This chapter is based on: Çelik, M. F., & Erten, E. (2022). Biophysical parameter estimation of crops from polarimetric synthetic aperture radar imagery with data-driven polynomial chaos expansion and global sensitivity analysis. Computers and Electronics in Agriculture, 194, 106781. https://doi.org/10.1016/j.compag.2022.106781

between biophysical parameters and only backscatter coefficients, there is also a large number of studies exploring the potential of polarimetric features for crops biophysical parameters retrieval [6]–[8]. [9], [10] and [11] revealed the relationship of backscatter coefficients with leaf area index (LAI) and biomass, while [12] revealed the relationship between backscatter coefficients and phenological stages. Over the past decade, most of the research in estimating the biophysical parameters has emphasized the use of not only the backscatter coefficients but also polarimetric features derived from SAR data. A variety of methods have been used to estimate the biophysical parameters. [13] and [14] used complex covariance matrices, characterizing the physical properties of crops based on polarimetric information, obtained from the time series of fully polarimetric SAR data for the estimation of crop phenology. Furthermore, in the study by [15], the phenological stage of crops was estimated by random forest using polarimetric features, instead of original polarimetric covariance matrix, and considered as a classification problem. On the other hand, [16] used the polarimetric feature derived from SAR data to create a regression model with random forest to estimate phenological stages. In [17], the dynamical models for the estimation of the phenological stage were carried out using the principal component analysis based approach. [18] used only the linear regression model to estimate the phenological stage. In addition to linear model, [19] used the polarimetric radar vegetation index for crop growth monitoring.

The study conducted by [20] explored the relationship between polarimetric features derived from dual-pol SAR images and their sensitivity to phenological stage of rice fields. It was indicated that although, dual-pol SAR images can be effective to estimate the phenological stages of rice, some phenological stages can be difficult to separate with each other. To overcome this problem, [21] emphasized that quad-pol SAR images resolve the relationship between phenology of rice fields and polarimetric features more accurate. In the study by [22], besides the phenological stages of crops, LAI and normalized difference vegetation index (NDVI), which are two important key parameters involved in a variety of agricultural studies, were estimated with a regression approach using polarimetric features derived from quad-pol SAR images. The non-linear regression problems were solved using orthogonal polynomials, and

their weights corresponding to the uncertainty of the estimation were used for the global sensitivity analysis (GSA). However, the interaction and correlation among polarimetric features, which were ignored, may undermine the usability of this high dimensional model representation for feature selection. Even if correlated input data do not significantly affect the regression result, the presence of the correlation among polarimetric features may provide unreliable results in the sensitivity analysis. In addition, the different orders of interaction between polarimetric features can be another factor affecting the regression result. In this study, one of the attractive metamodeling methods, namely polynomial chaos expansion (PCE), was considered to handle these two strong assumptions.

PCE is one of the most popular techniques, which establishes the functional relationship between the input data, including their interaction, and the output data using orthogonal polynomials, in uncertainty quantification, specifically in metamodeling for engineering [23]–[26]. Besides the frequent use of PCE in uncertainty quantification, studies show that the PCE can be used as a data-driven machine learning technique for regression analysis [27]–[29], and its direct relationship with sensitivity analysis based on orthogonal polynomial coefficients makes the model popular where there is a high dimensional input feature vector.

The sensitivity analysis is an efficient tool for identifying important input variables which drive the output. The sensitivity analysis is categorized into two groups, namely GSA and local sensitivity analysis. The local sensitivity analysis can determine the influence with partial derivatives at a given point which makes it insensitive to detect the dependency of the model (or output) based on the input variables. On the other hand, GSA can determine the influence of each input vector and their interaction with each other for the entire input spectrum rather than only at a local point. Concerning non-linear models, such as PCE, the interactions between the selected input variables are of great importance for the model. This response of the model cannot be identified by local sensitivity analysis [30].

PCE-based regression can be carried out using classical orthogonal polynomial functions depending on the distribution of the independent input variables [31]. In case

the independent variables are arbitrarily distributed, either arbitrary polynomials can be used, or the input variables can be transformed into another distribution (e.g. uniform space, standard normal, etc.) to be able to use classical orthogonal polynomials. The independence of the input variables allows us to perform the GSA using Sobol's method based on each input vector and their different interaction orders. The arbitrary polynomials can also be used together with the dependent input variables for the regression while preserving the original variable space, where the PCE is formulated based on the copula function to define the dependency structure of the input vector [28]. The variables can be made statistically independent by transforming them into a uniform space using Rosenblatt transformation [32] or into standard normal space using Nataf transformation [33]. Nevertheless, having an independent input has no significant effect on the accuracy of the regression analysis and shows similar accuracy with the regression analysis of dependent variables using arbitrary polynomials [27]. With Rosenblatt and Nataf transformations, since the vectors are transformed in an order depending on the cumulative density function of the first input vector, the Sobol Indices of the first input vector are always higher. Hence, as the order of the input vectors changes, different Sobol Indices emerge. This produces unreliable results in the GSA using Sobol's method for determining the significance level of input variables on the regression model. Though, the same problem may be observed even if the input variables are transformed into another space. Thus, every combination of the order of input variables must be tested to ultimately determine their level of significance to the output [34]. However, this is not feasible considering the amount of computational burden, specifically for a higher dimensional input feature vector.

In this study, we performed PCE to estimate biophysical parameters from polarimetric features generated from RADARSAT-2 data and in-situ measurements of LAI and NDVI, which were obtained within the AgriSAR2009 campaign [35]. The pre-processing step was applied to reduce the size of the generated polarimetric feature data that were statistically correlated even though some of them were physically independent. A principal component analysis based approach was implemented by [27] to reduce data size, and the effect of the interaction of input data in the PCE method on the regression was revealed. However, important polarimetric feature

selection analysis could not be performed since the original data was transformed into a new independent space. Therefore, in this study, Neighborhood Component Analysis (NCA) was applied to create a sub-feature space that is uncorrelated and important for the target parameter in the original data space. The effect of the obtained statistically independent polarimetric feature vectors on the biophysical parameters was then investigated. Finally, the importance of features that drive the regression models was investigated using GSA with Sobol's method.

This paper begins by laying out the theoretical dimensions of the research in Section 2.2, and it then goes on to the data set together with the pre-processing details used for this study in Section 2.3. Section 4.3 is concerned with the implementation details undertaken during the experimental study and presents the analysis findings, focusing on the two key themes: regression and GSA of biophysical parameters. We finalize the paper with discussions and comments about the study and some prospects in Section 2.5.

2.2 Methodology

2.2.1 Polynomial chaos expansion

PCE is an uncertainty quantification technique used to express Y as the function of $\boldsymbol{X}$ with the orthogonal polynomials:

$$Y = \mathscr{M}(\boldsymbol{X}) \tag{2.1}$$

The function $\mathscr{M}(\boldsymbol{X})$ is defined by multivariate orthogonal polynomial, which approximates Y as:

$$Y = \mathscr{M}(\boldsymbol{X}) \approx \sum_{\boldsymbol{i=1}}^{M} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}), \quad \boldsymbol{X} \overset{\text{def}}{=} \left\{ x^{(1)}, \ldots, x^{(N)} \right\} \tag{2.2}$$

where $\Psi_{\boldsymbol{\alpha}}$ is the multivariate orthogonal polynomial basis for $\boldsymbol{X}$, $y_{\boldsymbol{\alpha}}$ coefficient of the polynomials, N is the dimension of the input feature vector, and M is the truncation level [28].

The multi-dimensional orthogonal polynomial basis can be then constructed by product of univariate orthogonal polynomials if the input random vector elements are independent as:

$$\Psi_{\alpha}(\boldsymbol{x}) = \prod_{i=1}^{N} \phi_{\alpha_i}^{(i)}(x_i) \ , \tag{2.3}$$

where $\phi_{\alpha_i}^{(i)}$ is the univariate orthogonal polynomial with the truncate set of multi indices α_i. Note that, in reality the polynomial expansion in (2.2) is infinite and polynomial coefficients drastically increase for high dimensional feature vector $\boldsymbol{X}$ based on:

$$\begin{pmatrix} d+N \\ N \end{pmatrix}, \tag{2.4}$$

where d is the total polynomial order. For the practical purposes it truncates to a finite expansion M.

In addition to preventing the infinite expansion of the polynomial by degree, an additional truncation scheme can be settled to avoid the computational burden. Maximum interactions and hyperbolic (i.e., q-norm) parameters can be determined to prevent the increasing number of polynomial coefficients. While the maximum interaction parameter determines how many input vectors will interact, the q-norm parameter prevents higher-order polynomials from interacting with each other.

Finally, the unknown coefficients are estimated by minimizing the mean square residual error:

$$\hat{y} = \underset{y}{\operatorname{argmin}} \mathbb{E}\left[\left(y^{\top}\Psi(\boldsymbol{X}) - \mathscr{M}(\boldsymbol{X})\right)^2\right] + \lambda \|\boldsymbol{y}\|_1, \tag{2.5}$$

In this equation, the set of regression coefficients (y) that are formed based on the PCE are determined using least angle regression method [36]. The least angle regression method is an iterative linear regression technique similar to the classical forward stepwise regression method. It starts by initializing all the coefficients as zero. The algorithm moves the coefficients along the direction in a least-square sense to the predictor ($\Psi(\boldsymbol{X})$) with the highest correlation to the response value [37]. In each step, the leave-one-out (LOO) error is calculated for the residual value

$\left(\left(y^{\top}\Psi(\boldsymbol{X})-\mathscr{M}(\boldsymbol{X})\right)^2\right)$. Leave-one-out is a cross-validation method to overcome the over-fitting problem of the model. In this method, the metamodel that represents the output variable is computed for each i^{th} term where this term is left out from the computation of the metamodel and compared with the prediction value of the i^{th} term ($\mathscr{M}^{PC\setminus i}(x^{(i)})$.

$$\varepsilon_{LOO}=\frac{\sum_{i=1}^{s}\left(\mathscr{M}\left(x^{(i)}\right)-\mathscr{M}^{PC\setminus i}\left(x^{(i)}\right)\right)^2}{\sum_{i=1}^{s}\left(\mathscr{M}\left(x^{(i)}\right)-\hat{\mu}_Y\right)^2} \tag{2.6}$$

where s represents the number of samples and $\hat{\mu}_Y$ is the sample mean of the validation set.

At the end of iterations for the least angle regression method, the best set of sparse orthogonal polynomial basis is determined based on the minimum LOO error for the predictors. Detailed information about the theory of the implementation of PCE together with least angle regression methodology can be reached in [38].

2.2.2 Global sensitivity analysis

The sensitivity analysis is an efficient tool for identifying important input variables which drive the output. While the local sensitivity analysis can determine the influence at a given point with partial derivatives, GSA can determine the influence of each input vector and their interaction with each other for the entire input spectrum. Within the scope of this study, Sobol Indices were preferred for GSA to determine the effect of different order interactions by using PCE.

As Sobol stated, if $\mathscr{M}(\boldsymbol{x})$ given in Eq.(2.7) (a.k.a analysis of variance) is a square-integrable function and the input variables in uniform space [0,1] then left-handside of Eq.(2.8) is equal to 2^{nd} moment of $M(\boldsymbol{x})$.

$$\mathscr{M}(\boldsymbol{x})=\mathscr{M}_0+\sum_{i=1}^{M}\mathscr{M}_i(x_i)+\sum_{1\le i<j\le M}\mathscr{M}_{ij}\left(x_i,x_j\right)+\cdots+\mathscr{M}_{12\ldots M}(\boldsymbol{x}) \tag{2.7}$$

In this equation, $\mathscr{M}(\boldsymbol{x})$ is represented with the full set of coefficients on the right-hand side of the equation. It should be noted that in the practical implementation of Eq.

(2.7), the right-hand side is represented by the coefficients in the best set of sparse orthogonal polynomial basis (see Sect.2.2.1).

$$\int \left(\mathscr{M}^2(\boldsymbol{x}) - \mathscr{M}_0^2\right) = \int \sum_{i=1}^{M} \mathscr{M}_i^2\left(x_i\right) + \sum_{1\leq i<j\leq M} \mathscr{M}_{ij}^2\left(x_i, x_j\right) + \cdots + \mathscr{M}_{12\ldots M}^2(x) \quad (2.8)$$

So the sum of each term on the right-hand side of (Eq.2.9) gives the total variance.

$$D = \mathrm{Var}[\mathscr{M}^2(\boldsymbol{x})] = \int \sum_{i=1}^{M} D_i\left(x_i\right) + \sum_{1\leq i<j\leq M} D_{ij}\left(x_i, x_j\right) + \cdots + D_{12\ldots M}(x) \quad (2.9)$$

The ratio of each term in Eq.(2.9) to D allows us to calculate the global sensitivity indices (a.k.a Sobol Indices)

$$S_{i_1,\ldots,i_s} = D_{i_1,\ldots,i_s}/D \quad (2.10)$$

The sum of the variance of each variable is equal to 1 (Eq.2.11).

$$\sum_{i=1}^{n} S_i + \sum_{1\leqslant i<j\leqslant n} S_{ij} + \cdots + S_{1,2,\ldots,n} = 1 \quad (2.11)$$

While the term S_i means the individual effect of the input variable or 1^{st} order Sobol indices, $S_{i,j}$ and $S_{1,2,\ldots,n}$ show the effect of the double and higher-order interaction of the variables, respectively. The sum of terms with S containing any variable i gives the total Sobol indices of that variable.

2.3 Data Set Explanation

The data set obtained during the AgriSAR2009 campaign funded by ESA was used in this study. The campaign was carried out in three different regions. The majority of the data was acquired from Indian Head in Canada, followed by Flevoland in Holland and Barrax in Spain [35]. Only the data set collected in the Indian Head region was evaluated in this research, which covers eight fields with four crop types: field pea, barley, canola, and oat (see Figure 4.1).

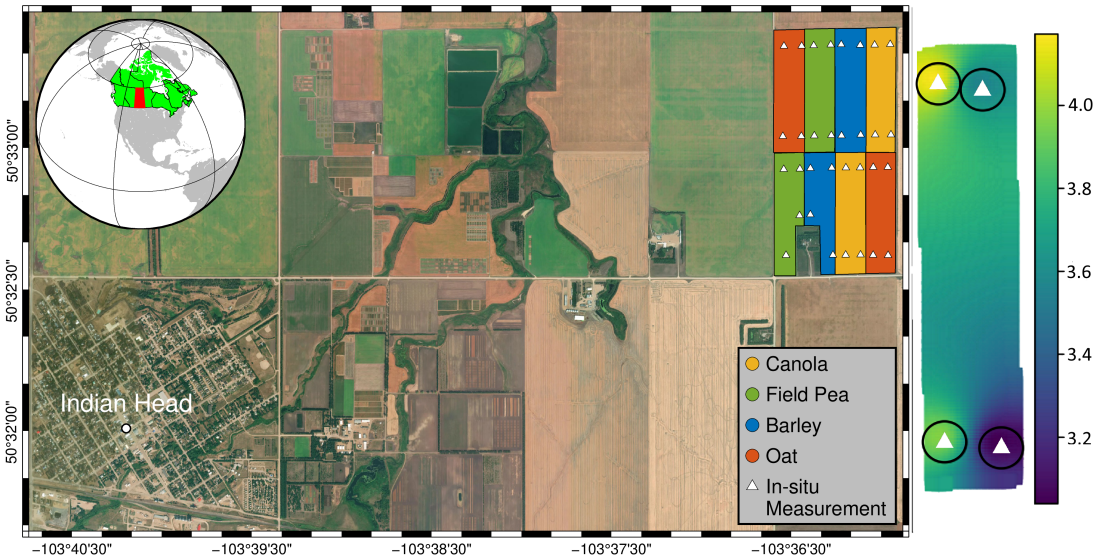


Figure 2.1 : Agricultural fields located in Indian Head (left), as well as the LAI representation of in-situ sampling set up (right).

In a single growth period covering June to July, fully polarimetric RADARSAT-2 images were acquired, and the in-situ measurements were conducted during the corresponding growing cycle [13,17,22,39]. As shown in Figure 4.1, field data, including LAI, NDVI, and phenological stage in BBCH scale (Biologische Bundesanstalt, Bundessortenamt and CHemical industry), were collected in the four corners of each field. In order to determine the output value corresponding to each pixel, the inverse distance weighting interpolation method was applied to the NDVI and LAI parameter. For controlling the reliability of the interpolated in-situ measurements in each field, pixels that are within the 7-pixel circular radius from the measurement location were used in the regression analysis. The variation of pixel number is related to the in-situ measurement location because some of them are near the edge of field which causes a loss of pixels. There are 120 to 145 pixels for each in-situ measurement location. The total number of pixels for each field is given in Table 2.1.

Table 2.1 : The number of samples per crop field.

Crop Type	Field ID	Pixels
Barley	PF-2	3960
Barley	PF-7	4264
Field Pea	PF-1	4136
Field Pea	PF-6	4392
Canola	PF-3	4400
Canola	PF-8	4584
Oat	PF-4	4136
Oat	PF-5	4344

Fully polarimetric RADARSAT-2 images and in-situ measurements had different temporal frequencies, which means that the over-pass time of the SAR images did not coincide with the in-situ measurements [39]. In order to minimize the effect of the temporal inconsistencies, data with more than three days time difference between the in-situ measurements and the polarimetric synthetic aperture radar (PolSAR) images were ignored. After data cleaning, 8 RADARSAT-2 images (see details in Table2.2)and their corresponding in-situ measurements exist in the phenological growth period of crops. It is also worth to noting that the impact of the incidence angle is ignored for not decreasing the number of independent samples.

Table 2.2 : The RADARSAT-2 images information that used in the study [35].

Acqusition Date	Pass	Incidence Angle
6/3/2009	Descending	22
6/11/2009	Ascending	35
6/17/2009	Descending	30
6/24/2009	Descending	34
7/1/2009	Descending	39
7/11/2009	Descending	30
7/12/2009	Ascending	31
7/26/2009	Ascending	22

RADARSAT-2 images cover the growth period from seeding until the end of the vegetative stage for canola. However, for cereals and field pea, the images cover only the period between seeding to reproductive stage. Total 23 polarimetric features (see Figure 2.3) created from the fully polarimetric images were then used to build a regression model to identify the NDVI and LAI parameters within the test samples.

All the features were computed after a multi-looking with a moving-average window of 9×9 pixels.

Table 2.3 : Polarimetric features derived from full polarimetric SAR images.

Feature Index	Feature Description	Feature Symbol
F1	Span of the Covariance Matrix*	SPAN
F2, F3, F4	Backscatter Coefficients*	$\|HH\|^2, \|HV\|^2, \|VV\|^2$
F5, F6, F7	Eigenvalue/vector Decomposition Elements	H, A, α
F8, F9, F10	Backscatter Coefficients Ratios*	$\|HH\|/\|VV\|$, $\|HV\|/\|VV\|$, $\|HV\|/\|HH\|$
F11, F12, F13	Correlation Between Channels	$\rho_{HH,VV}, \rho_{HH,HV}, \rho_{VV,HV}$
F14, F15, F16	Phase Differences Correlation	$\phi_{HH,VV}, \phi_{HH,HV}, \phi_{VV,HV}$
F17, F18	1.- 2. Pauli Components	$\|HH+VV\|^2, \|HH-VV\|^2$
F19	1.- 2. Pauli Components Phase Differences	$\phi_{HH+VV,HH-VV}$
F20	1.- 2. Pauli Components Correlation	$\rho_{HH+VV,HH-VV}$
F21, F22, F23	Shannon Entropy - Intensity, Polarization and Dual	SEI, SEP, SED

* backscatter coefficients are in decibel scale

In order to understand the impact of sampling on the model, two different cases were examined for building regression models:

- Case I: Each crop type and field evaluated separately,
- Case II: Each crop type evaluated within itself.

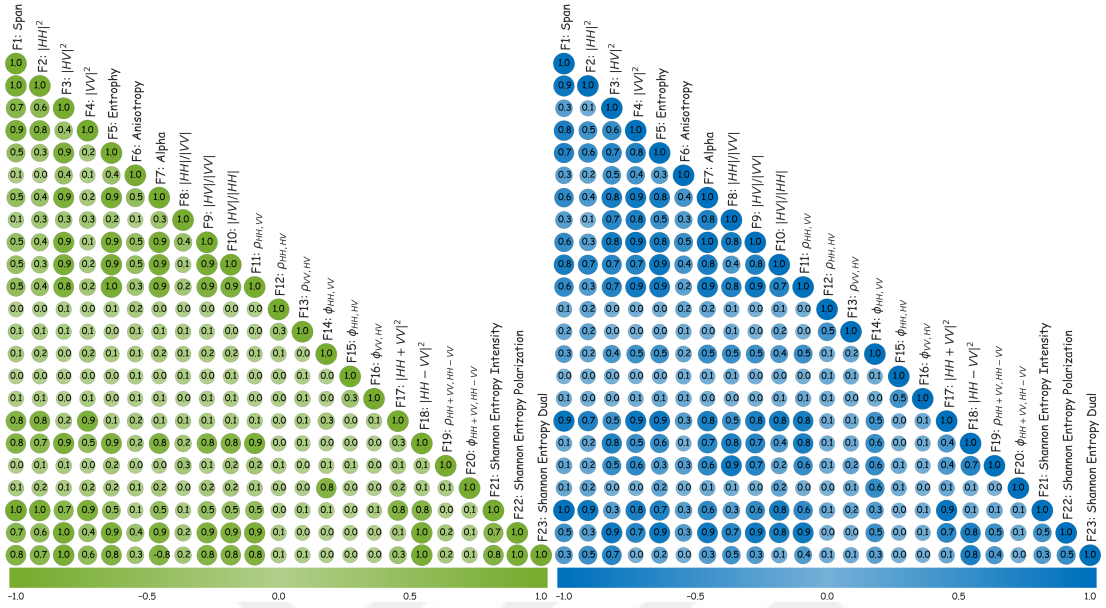
After this section, green, blue, yellow, and orange colors will be used for field pea, barley, canola, and oat crops, respectively, to better understand the figures.

2.4 Experimental Study and Discussions

2.4.1 Pre-processing: PCE settings

A close look at the input vector $\boldsymbol{X}$ in $Y = \mathscr{M}(\boldsymbol{X})$ given in Figure 2.2 shows the presence of high correlations among the PolSAR features, which have strong effects on the regression and similarly on Sobol indices. In order to perform regression on the basis, the input vectors should be statistically independent. One approach could be ignoring the correlations between random inputs and calculating the basis, which is the case for most of the studies in the uncertainty propagation [22,28]. Even though a basis of polynomial orthonormals can be built and regression can be performed, this approach is too costly and limits the usage of the Sobol indices for the identification of the most

influential input variables. In this paper, NCA approach [40] was performed for the subset selection from all polarimetric features to form an uncorrelated low dimensional feature space, instead of the PCA approach carried out by [27].



(a) Field Pea (b) Barley

(c) Canola (d) Oat

Figure 2.2 : Input feature vector correlations for each crop type. The polarimetric features are chosen according to their extensive usage in crop monitoring studies [15,17,18,22]. In the figure, $\rho_{i,j}$ and $\phi_{i,j}$ show the degree of correlation and the phase difference between two polarimetric channels, i and j, respectively.

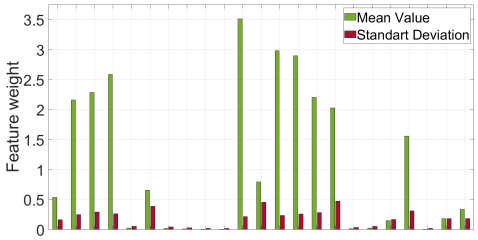
To find a general feature subset for crop types, all pixel values that belong to each crop type are evaluated separately, like in Case-II, and features that are found important are used for both cases (Case-I and Case-II). The feature subset selection was then applied in two steps. In the first step, NCA was carried out to calculate the feature ranking. In order to ensure the stability of the feature ranking, the NCA analysis was applied 50 times, and elimination was made according to the mean value of the ranking of the feature. A threshold (t) value was determined to eliminate features below $t = 2$ for the LAI parameter and the value of $t = 1.5$ for the NDVI parameter. In the second step of the feature selection, remaining feature correlations were checked according to the Spearman correlation coefficient, in case there is a monotonic relationship between features. If the correlation value between features was greater than 0.5 in absolute terms, the data with the lower value in the feature ranking was also eliminated. The important, informative, and uncorrelated features for different crop types according to NDVI and LAI parameters were given in the Table 2.4.

Table 2.4 : Selected polarimetric features for regression

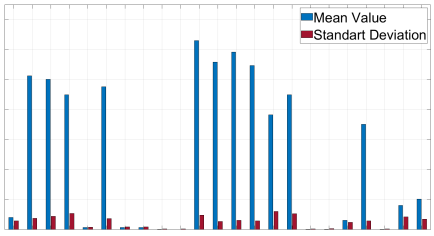
Crop Type	NDVI Features	LAI Features
Barley	$\|HH\|^2, \|VV\|^2$	$\|HH\|^2, A$
	$\phi_{HH,HV}, \phi_{VV,HV}, \rho_{HH+VV,HH-VV}$	$\rho_{HH,VV}, \rho_{HH,HV}, \rho_{VV,HV}, \phi_{VV,HV}$
Field Pea	$\|HV\|^2, \|VV\|^2$	$\|VV\|^2, \rho_{HH,VV}, \rho_{VV,HV}$
	$\phi_{HH,HV}, \phi_{VV,HV}, \rho_{HH+VV,HH-VV}$	$\phi_{HH,VV}, \phi_{HH,HV}, \phi_{VV,HV}$
Canola	$\|HV\|^2, \|VV\|^2$	$\|HH\|^2, \|HV\|^2$
	$\phi_{HH,HV}, \phi_{VV,HV}, \rho_{HH+VV,HH-VV}$	$\phi_{HH,HV}, \rho_{HH+VV,HH-VV}$
Oat	$\|HH\|^2, \|HV\|^2$	$\|HH\|^2, \|HV\|^2$
	$\phi_{HH,HV}, \phi_{VV,HV}, \rho_{HH+VV,HH-VV}$	$\rho_{VV,HV}, \phi_{HH,VV}, \phi_{VV,HV}$

Different features were found significant for LAI (Figure 2.3) and NDVI (Figure 2.4) in different significance levels. While the same six PolSAR features were important for the NDVI parameter according to feature ranking, there are slight differences for the LAI parameter. Although NDVI and LAI are key determinants of crop growth, and their tight relationship has been studied densely in remote sensing, their dynamics are not linearly related. NDVI, as a measure of vegetation greenness, is much less sensitive to the morphology of crops, such as leaf shape, leaf density, moisture, etc. This explains the heterogeneity of important features for different crops in the case of LAI estimation. PolSAR features related to the cross polarizations are highly sensitive

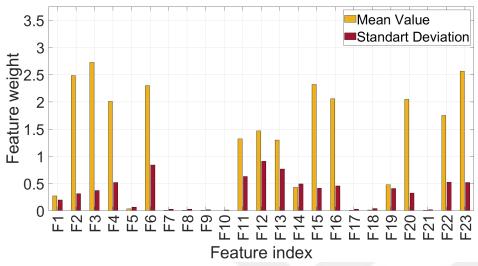
to in-situ NDVI for all crop types. The important features for LAI differ slightly among the crops. However, some common features are still important, which are related to the randomness of vegetation and leaf morphology.



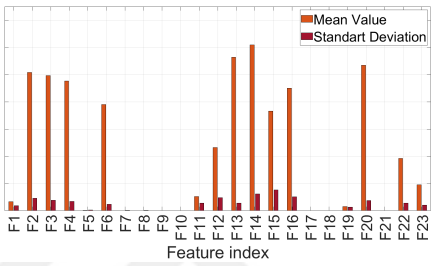
(a) Field Pea



(b) Barley

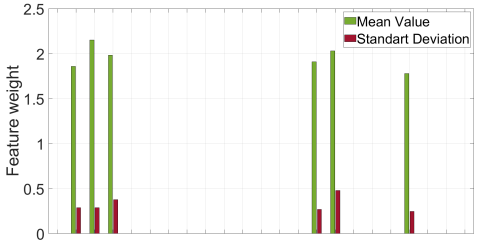


(c) Canola

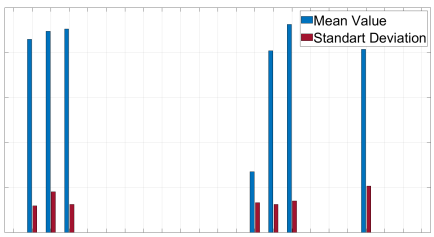


(d) Oat

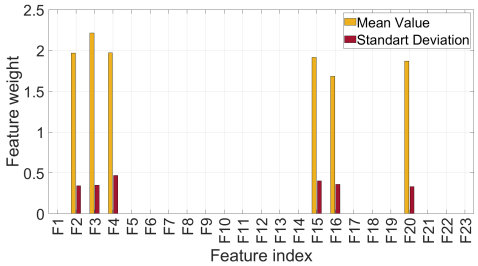
Figure 2.3 : Feature ranking for LAI parameter based on crop type.



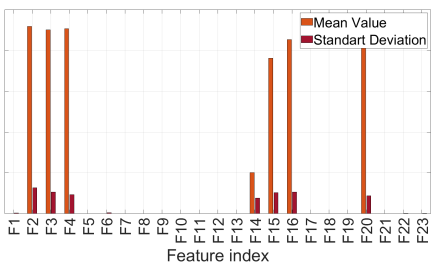
(a) Field Pea



(b) Barley



(c) Canola



(d) Oat

Figure 2.4 : Feature ranking for NDVI parameter based on crop type.

As stated in [36], higher-order interactions between input vectors and higher-degree interactions of the polynomial have no significant effect on the model. Within the scope of this study, maximum triple interaction and q-norm (q=0.75) were determined as the truncation scheme of the interactions. The polynomials are expanded from the 3^{rd} to the 25^{th} degree to determine the optimum expansion degree of the polynomial. Input vectors are transformed into uniform space between -1 and 1 using iso-probabilistic transformation to ensure the orthogonality of the Legendre polynomial. Each iteration is repeated 50 times with randomly selected input samples. In Figure 2.5, it can be seen that expanding the orthogonal polynomial beyond 10^{th} degree does not change the estimation accuracy significantly. Besides, increasing the expansion degree of the polynomial brings extra effort to the computation.

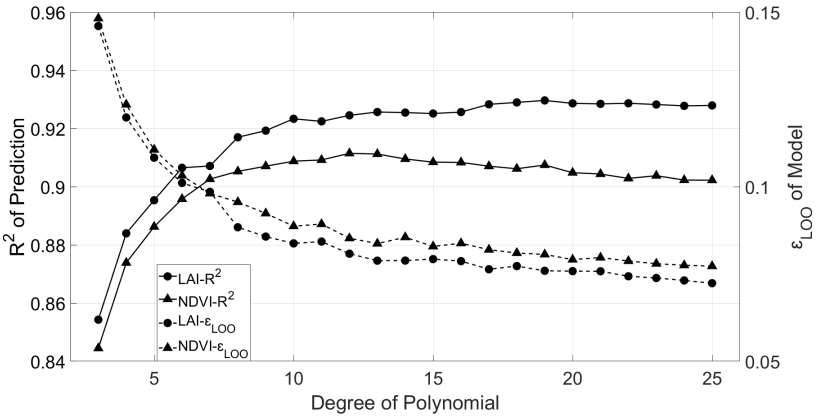


Figure 2.5 : Accuracy analysis of the estimation of LAI&NDVI with respect to polynomial degree.

2.4.2 PCE processing

The PCE regression with original features can be implemented when the independence of the features is assured. Before assessing the contribution of PolSAR features for biophysical parameter estimation using PCE, the multi-dimensional orthogonal polynomials should be constructed with the selected statistically independent features (see Table 2.4). For each crop type and sampling strategy (Case-I and Case-II), the pixel values were divided into training (25%) and testing (75%) samples to conduct the data-driven PCE regression on two different biophysical parameters, NDVI and LAI. Leave-one-out error measurement was determined to prevent over-fitting of the

regression model and to ensure the training accuracy. The root mean square error (RMSE) and adjusted R^2 were used to evaluate the performance of PCE regression for predicting the biophysical parameters.

As mentioned in the previous section 2.4.1, the feature ranking was determined for each crop type as described in Case-II and used in the regression analysis for each biophysical parameter for Case-I and Case-II. For both cases, the method of PCE was carried out by considering the maximum triple interaction up to 10^{th} degree polynomials. Each regression analysis was calculated with 50 runs in which the training and the testing data in the data set are randomly selected, and the mean and the standard deviation values of 50 runs were given in Table 2.5 and Table 2.6. According to adjusted R^2 and *RMSE* results, we can say that Case-I and Case-II sampling approaches do not significantly affect the regression accuracy. In addition, based on the regression results, the accuracy of the NDVI parameter was slightly higher than that of the LAI parameter. One of the possible reasons for this is that the LAI parameter is directly related to the physical properties of the crops. Instead, NDVI is more related to chlorophyll content, making the element of the input feature vector more homogeneous for the entire field. Due to this aspect, the LAI parameter is more sensitive for the polarimetric features derived from the SAR data.

Table 2.5 : Accuracy analysis of PCE for LAI.

Crop Type	Field ID	LOOe	Adj. R^2	RMSE
Case I				
Field Pea	PF-1	0.09±0.006	0.91±0.005	0.38±0.011
	PF-6	0.07±0.005	0.94±0.004	0.28±0.009
Barley	PF-2	0.05±0.004	0.95±0.006	0.37±0.020
	PF-7	0.08±0.008	0.92±0.007	0.46±0.019
Canola	PF-3	0.07±0.003	0.93±0.006	0.36±0.017
	PF-8	0.04±0.003	0.96±0.005	0.26±0.015
Oat	PF-4	0.06±0.004	0.95±0.005	0.22±0.012
	PF-5	0.08±0.005	0.93±0.005	0.27±0.009
Case II				
Field Pea	PF-1,6	0.10±0.004	0.90±0.004	0.37±0.007
Barley	PF-2,7	0.09±0.004	0.91±0.003	0.48±0.009
Canola	PF-3,8	0.06±0.003	0.94±0.004	0.33±0.010
Oat	PF-4,5	0.08±0.003	0.93±0.003	0.28±0.006

Table 2.6 : Accuracy analysis of PCE for NDVI.

Crop Type	Field ID	LOOe	Adj. R^2	RMSE
Case I				
Field Pea	PF-1	0.02±0.030	0.98±0.002	0.03±0.003
	PF-6	0.02±0.028	0.98±0.001	0.03±0.002
Barley	PF-2	0.10±0.082	0.90±0.009	0.08±0.006
	PF-7	0.06±0.063	0.94±0.005	0.06±0.005
Canola	PF-3	0.02±0.032	0.98±0.001	0.03±0.004
	PF-8	0.01±0.024	0.99±0.001	0.02±0.002
Oat	PF-4	0.05±0.050	0.96±0.004	0.05±0.004
	PF-5	0.07±0.062	0.94±0.004	0.06±0.004
Case II				
Field Pea	PF-1,6	0.02±0.030	0.98±0.001	0.03±0.002
Barley	PF-2,7	0.08±0.073	0.92±0.004	0.07±0.003
Canola	PF-3,8	0.02±0.030	0.98±0.001	0.03±0.002
Oat	PF-4,5	0.07±0.062	0.93±0.003	0.06±0.003

2.4.3 PCE post-processing: GSA for LAI and NDVI

Following the regression, Sobol sensitivity analysis was performed to understand not only the major drivers of the variation of the biophysical parameters, but also the importance of their interactions on the regression model. Aside from the fact that the degree of the polynomials beyond 10 have no significant contribution to the regression analysis, higher degree polynomials drastically increase the computational effort (see Figure 2.5). On the other hand, PCE coefficients can be directly used to calculate all orders of Sobol Indices without additional computational burden. The Sobol Indices show the significance of each feature and its interaction with the estimated parameter. Even though the magnitude of the 1^{st} order Sobol Indices are quite high, 2^{nd} and 3^{rd} order interactions have an effect on the regression and should not be excluded from the analysis.

The GSA was applied with the calculated polynomial coefficients after reaching the targeted accuracy criteria in the regression analysis using PCE. The effect of each polarimetric feature on the regression and its interaction with other features were determined with GSA as in Eq.(2.10). Table 2.7 and Table 2.8 shows the result of GSA for LAI and NDVI, respectively. It can be seen from the tables that double and triple interactions are higher in the regression of the LAI parameter than in the

regression of the NDVI parameter. While triple interactions are smaller for estimating the NDVI parameter compared to the LAI parameter, they are even negligible for field pea and canola. On the other hand, the 3^{rd} order interactions are more significant in the regression of the LAI parameter.

Table 2.7 : Sobol Sensitivity Analysis for LAI.

Crop Type	Field ID	1^{st} Order	2^{nd} Order	3^{rd} Order
Case I				
Field Pea	PF-1	0.77 ± 0.04	0.16 ± 0.03	0.07 ± 0.02
	PF-6	0.72 ± 0.04	0.21 ± 0.04	0.07 ± 0.01
Barley	PF-2	0.84 ± 0.02	0.12 ± 0.02	0.04 ± 0.01
	PF-7	0.80 ± 0.03	0.12 ± 0.02	0.08 ± 0.02
Canola	PF-3	0.76 ± 0.03	0.19 ± 0.02	0.04 ± 0.01
	PF-8	0.58 ± 0.06	0.37 ± 0.06	0.05 ± 0.01
Oat	PF-4	0.77 ± 0.02	0.14 ± 0.02	0.09 ± 0.01
	PF-5	0.77 ± 0.03	0.11 ± 0.02	0.12 ± 0.02
Case II				
Field Pea	PF-1,6	0.72 ± 0.04	0.21 ± 0.04	0.07 ± 0.01
Barley	PF-2,7	0.84 ± 0.02	0.11 ± 0.01	0.05 ± 0.01
Canola	PF-3,8	0.75 ± 0.02	0.21 ± 0.02	0.04 ± 0.01
Oat	PF-4,5	0.77 ± 0.02	0.15 ± 0.01	0.08 ± 0.01

Table 2.8 : Sobol Sensitivity Analysis for NDVI.

Crop Type	Field ID	1^{st} Order	2^{nd} Order	3^{rd} Order
Case I				
Field Pea	PF-1	0.93 ± 0.01	0.06 ± 0.01	0.01 ± 0.00
	PF-6	0.93 ± 0.01	0.06 ± 0.01	0.01 ± 0.00
Barley	PF-2	0.75 ± 0.03	0.22 ± 0.03	0.03 ± 0.02
	PF-7	0.80 ± 0.03	0.16 ± 0.03	0.04 ± 0.01
Canola	PF-3	0.96 ± 0.00	0.03 ± 0.00	0.01 ± 0.00
	PF-8	0.94 ± 0.01	0.05 ± 0.01	0.01 ± 0.00
Oat	PF-4	0.72 ± 0.03	0.22 ± 0.03	0.05 ± 0.01
	PF-5	0.87 ± 0.01	0.09 ± 0.01	0.04 ± 0.01
Case II				
Field Pea	PF-1,6	0.93 ± 0.01	0.06 ± 0.01	0.01 ± 0.00
Barley	PF-2,7	0.76 ± 0.03	0.21 ± 0.03	0.03 ± 0.01
Canola	PF-3,8	0.95 ± 0.00	0.04 ± 0.00	0.01 ± 0.00
Oat	PF-4,5	0.81 ± 0.02	0.14 ± 0.02	0.05 ± 0.01

Figure 2.6, on the left side, presents the mean value and standard deviation of total Sobol indices for each feature that was found significant. The total Sobol indices

of each feature include maximum triple interactions. The right side of Figure 2.6 presents the accuracy of the regression according to the Total Sobol Indices in terms of both mean value of adjusted R^2 and RMSE of 50 runs. It should be noted that the accuracy plots begin with the individual effect of the feature with the highest total Sobol indices on the regression model and add the rest of the features based on their total Sobol indices sequentially, including their interactions, and its interactions are truncated maximum 3^{rd} order. To give an example, the accuracy plot, shown in Figure 2.6(b), starts with the individual effect of feature $F4 : |VV|^2$, which corresponds to the value with the highest Total Sobol Indices in Figure 2.6(a). The feature $F11 : \rho_{HH,VV}$ is, later, included in the analysis to clarify the effect of $F4 : |VV|^2$ and $F11 : \rho_{HH,VV}$ features on the regression, together with their double interactions. After the addition of $F15 : \phi_{HH,HV}$ and the rest of the features, in an order based on their total Sobol indices, the individual effect of each feature (1^{st} order) and their double and triple interactions are considered in the analysis.

The interpretation of the accuracy plots for both LAI and NDVI parameters are made with respect to the adjusted R^2 values. Nevertheless, in a reverse sense, the same behavior can be observed for the RMSE values as well.

For the regression of the LAI parameter of field pea, the Total Sobol Indices of polarimetric feature $F4 : |VV|^2$ is slightly higher than that of $F11 : \rho_{HH,VV}$. When regression analysis is performed with $F4 : |VV|^2$ alone, the estimation accuracy is around $R^2 = 0.35$. When polarimetric feature $F11 : \rho_{HH,VV}$ is added to the regression together with its double interaction with $F4 : |VV|^2$, the prediction accuracy reaches around $R^2 = 0.83$. By adding the rest of the polarimetric features sequentially to the regression model, including their double and triple interactions, the prediction accuracy slowly approaches $R^2 = 0.90$.

It can be clearly seen in Figure 2.6(c) that polarimetric feature $F11 : \rho_{HH,VV}$ dominates the regression of LAI parameter for barley. The prediction accuracy reaches 0.85 when the regression is carried out with only $F11 : \rho_{HH,VV}$. When adding other polarimetric features together with their interactions, the prediction accuracy increases up to $R^2 = 0.91$.

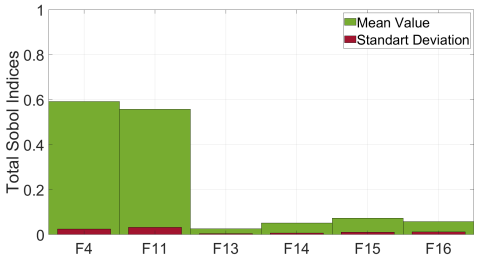
The estimation accuracy obtained by the regression of the LAI parameter with a single polarimetric feature in the canola is 0.85, as it was for barley. However, unlike barley, the polarimetric feature $F3 : |HV|^2$ influences the regression model the most for canola. When polarimetric feature $F2 : |HH|^2$ is added, the estimation accuracy of the regression exceeds $R^2 = 0.90$. After other parameters are added in order of importance, the estimation accuracy reaches $R^2 = 0.95$ at most.

Oat follows a similar trend with the canola. $F3 : |HV|^2$ parameter is more important than other parameters for the regression of the LAI parameter of canola. However, the total Sobol indices of $F3 : |HV|^2$ parameter is lower for oat than that of canola. Therefore, the estimation accuracy is approximately $R^2 = 0.65$ by regression of the $F3 : |HV|^2$ parameter alone for LAI parameter of oat, compared to the R^2 value of 0.85 for canola for the same feature. The distribution of the significance of features for oat has different characteristics compared to the rest of the crop types. While the regression model of the LAI parameter for other crop types is usually driven by one or a maximum of two features, the estimation accuracy for oat, after adding the two features that are found more important, continues to increase gradually when four features are included with their interactions. The prediction accuracy reaches $R^2 = 0.75$ with the addition of the parameter $F2 : |HH|^2$, which is of secondary importance, $R^2 = 0.85$ with the parameter $F14 : \phi_{HH,VV}$, which has the third level of importance. After adding the feature $F20 : \rho_{HH+VV,HH-VV}$, the accuracy becomes $R^2 = 0.90$. Using the last two polarimetric feature data increases the estimation accuracy of the regression from $R^2 = 0.90$ to $R^2 = 0.93$.

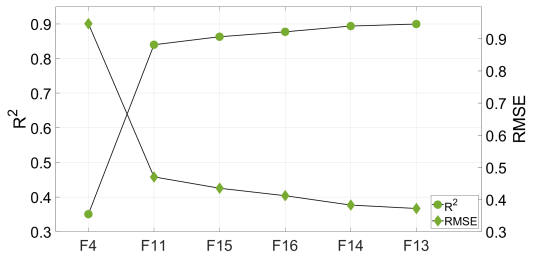
Unlike the regression analysis of the LAI parameter, it was found that in the regression of the NDVI parameter, there is only one polarimetric feature that is much more effective than the other polarimetric features for the regression model of each crop type. The total Sobol indices of the polarimetric feature $F3 : |HV|^2$, which drives the regression model for field pea, canola, and oats, are approximately 1, 0.90, and 1, respectively. The regression model of barley, on the other hand, is dominated by polarimetric feature $F4 : |VV|^2$. When the regression is implemented with the dominating feature, the estimation accuracy is approximately $R^2 = 0.94$, $R^2 = 0.88$, $R^2 = 0.94$, and $R^2 = 0.76$ for field pea, barley, canola, and oat, respectively.

Correspondingly, the estimation accuracy of all crop types is higher than the LAI parameter estimation using only the feature with the highest total Sobol indices. It is obvious that the backscatter coefficients are dominant for the estimation of the NDVI parameter. When the second parameter for field pea and canola is added, the estimation accuracy of the regression results reaches the value of $R^2 = 0.97$. With the addition of other parameters, the estimation accuracy is approximately $R^2 = 0.99$. The prediction accuracy reaches $R^2 = 0.90$ when the top three most important polarimetric features for barley and oat and their interactions are included in the regression. In general, the regression of the NDVI parameter of oat and barley yields lower accuracy than the other two crop types. It should be noted that the 3^{rd} order Sobol indices of field pea and canola are found insignificant for the regression model of NDVI parameter. The corresponding values of barley and oat are noticeably smaller compared to the 1^{st} and 2^{nd} order Sobol indices.

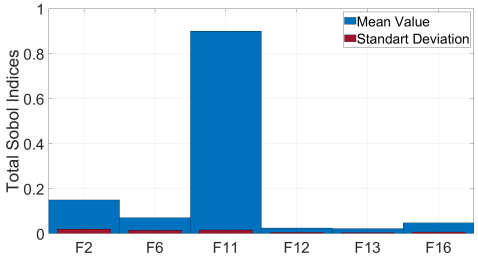
Together these results provide important insights into the effect of interactions between polarimetric features on the regression model.



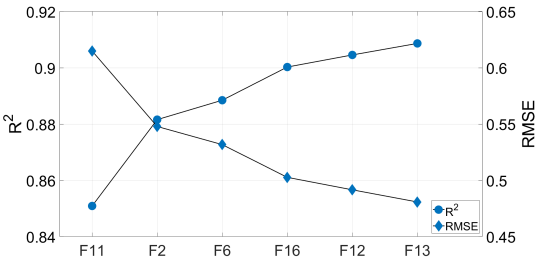
(a) Field pea



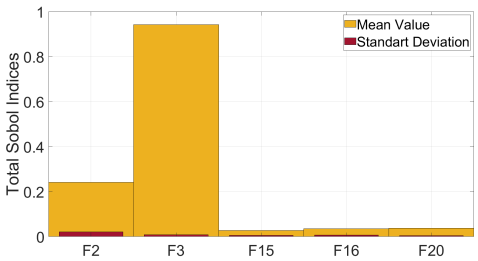
(b) Features effect on regression analysis for field pea



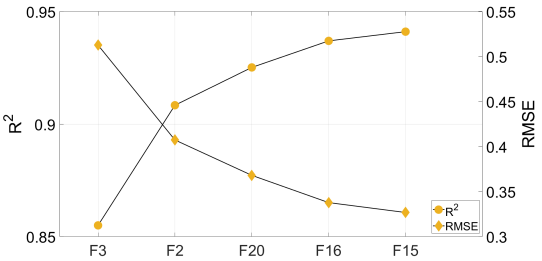
(c) Barley



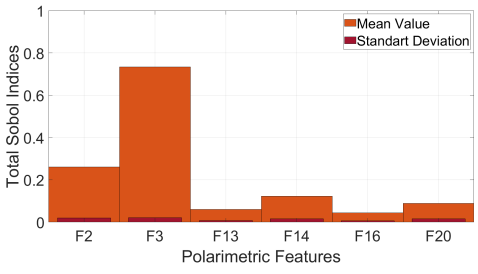
(d) Features effect on regression analysis for barley



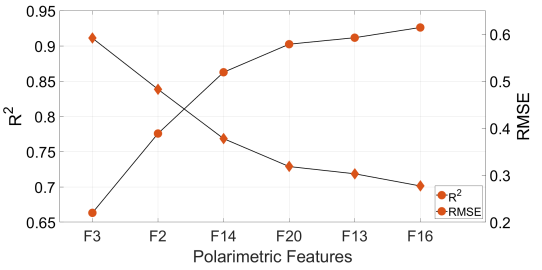
(e) Canola



(f) Features effect on regression analysis for canola

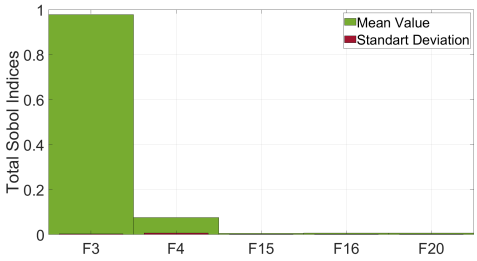


(g) Oat

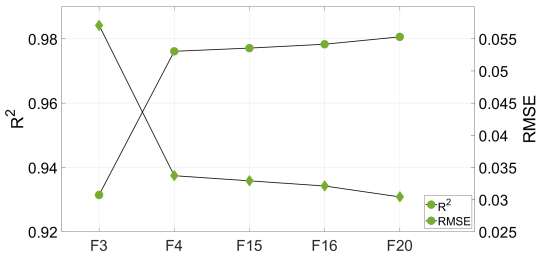


(h) Features effect on regression analysis for oat

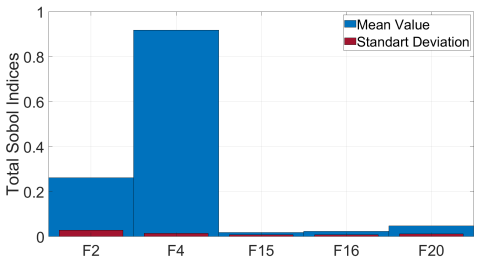
Figure 2.6 : Total Sobol indices for each crop's LAI (on the left) given with their corresponding regression accuracy (on the right) after selecting the features based on the total Sobol indices, showing how the number of ranked features at each run of the prediction affects the prediction results.



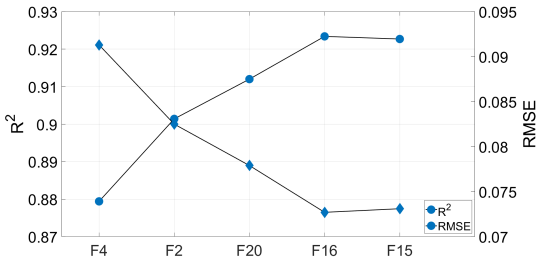
(a) Field pea



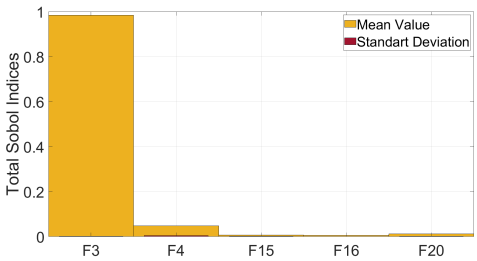
(b) Features effect on regression analysis for field pea



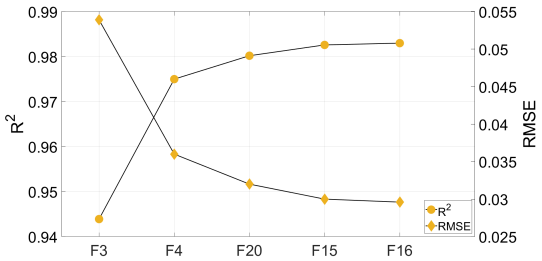
(c) Barley



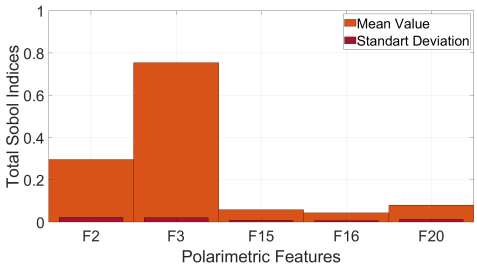
(d) Features effect on regression analysis for barley



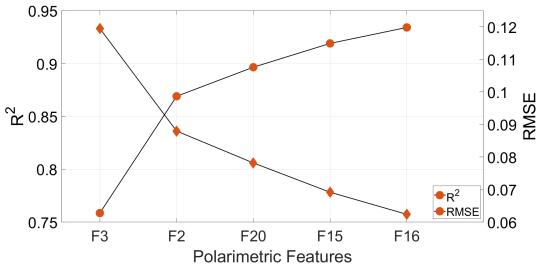
(e) Canola



(f) Features effect on regression analysis for canola



(g) Oat



(h) Features effect on regression analysis for oat

Figure 2.7 : Total Sobol indices for each crop's NDVI (on the left) given with their corresponding regression accuracy (on the right) after selecting the features based on the total Sobol indices, showing how the number of ranked features at each run of the prediction affects the prediction results.

2.5 Conclusion and Future Work

PCE is a machine learning method that can be implemented for the estimation of biophysical parameters. The GSA can be directly carried out from the coefficients of orthonormal basis. Once the independent input features are set, only three parameters have to be optimized: the polynomial degree, interactions (double, triple, or higher), and their level (q-norm). All orders of Sobol Indices can be calculated to eliminate or understand insignificant features and their interactions between them based on the PCE regression model. The 1^{st} order Sobol indices give the individual effect of each input vector, while the second and third Sobol indices give the effect of the interaction of the input vectors. Total Sobol Indices provide information about the individual effect of the features and the effect of their interactions on regression.

The work presented in this paper assessed the usage of data-driven PCE and its direct usage in GSA for biophysical parameter estimation. The PCE-based regression analysis is performed for LAI and NDVI parameters for the agricultural fields located on the east side of the Indian Head region of Saskatchewan, Canada. The data set used in this study belongs to the largest agriculture field out of three test areas given in the AgriSAR2009 campaign by ESA. The data set includes in-situ measurements of LAI and NDVI parameters collected during the growth cycle of four crop types: field pea, barley, canola, and oat. The PolSAR features calculated from fully polarimetric RADARSAT-2 images were used to estimate the LAI and NDVI parameters. The NCA was applied to PolSAR features to have statistically independent features to reduce the number of input variables and the computational effort. Based on the threshold value for the correlation of the features, different sub-feature sets were created for each crop type. Using the sub-feature set for each crop, stochastic PCE regression models were built for GSA to understand the importance level of PolSAR features that were found significant based on NCA, together with their interactions (maximum triple interactions) for different biophysical parameters. In GSA, the total Sobol indices for the polarimetric features were computed to show their significance based on their effect on the regression model. The highest value of Total Sobol Indice shows the feature that dominates the estimation of the biophysical parameter.

The PCE results showed that beyond the 10^{th} degree of polynomial expansion does not contribute significantly to the estimation result and creates over-fitting the regression model while increasing the computational effort. Based on the GSA results, it can be concluded that the regression model of the NDVI parameter is primarily dominated by the polarimetric features related to the backscatter coefficients. However, this was not the case for the estimation of the LAI parameter. In addition to the backscatter coefficient, $F11 : \rho_{HH,VV}$ has been found effective for field pea and barley crops while estimating the LAI parameter. These results are in line with the studies by [39], [17] and [22] on the same data set, even though the approaches are different. Regression models were built for each crop field separately for the first case (Case-I) to clarify the impact of different sampling strategies. In the second case (Case-II), each crop type was evaluated within itself. The regression accuracy of both cases showed that there is no point in evaluating each field separately.

The present study was designed to determine the effect of polarimetric features and their interactions on the PCE regression model. In this context, PCE regression analysis with independent sub-feature space and GSA provided useful information about the importance of features on successfully estimating the biophysical parameters of crops. It should be mentioned that in order to enlarge the analysis conducted with PCE, one may interpolate the LAI and NDVI measurements to the acquisition date of RADARSAT-2 images to be able to include the images that were removed because of the temporal inconsistencies between the measurement dates of LAI and NDVI and the overpass time of the satellite, hence the effect of incidence angles of images onto biophysical parameters can be investigated.

This study may serve as the basis for regression-based biophysical parameter estimation in agricultural studies, specifically with increasingly large amounts of freely available remote sensing data. This would improve the estimation problem with dense in-time data to comprehensively clarify the importance of polarimetric features. In this context, Sentinel-1 has considerable potential in regression-based biophysical parameter estimation studies thanks to its temporal resolution and its polarimetric features highlighted in this study. Moreover, dense in-time Sentinel-2 images could be used to obtain estimated LAI and NDVI measurements, and unsupervised domain

adaptation between these two satellite's features could be operational in biophysical parameter estimation.

3. SOIL MOISTURE PREDICTION FROM REMOTE SENSING IMAGES COUPLED WITH CLIMATE, SOIL TEXTURE AND TOPOGRAPHY VIA DEEP LEARNING[1]

3.1 Introduction

Fresh water resources are depleting daily due to climate change and the increasing world population. Hence, effective use of available water is of utmost importance, which makes its monitoring vital for water savings, mitigation, and adaptation to climate change. In the last decade, soil moisture (SM) monitoring has been investigated with its different aspects, covering drought monitoring [41,42], flood prediction [43] and agricultural applications [44,45]. Particularly in agriculture, SM significantly impacts planning, seeding, fertilization, and irrigation activities. Besides, its close relationship with crop productivity makes SM monitoring an essential factor for optimizing the use of available water resources [46,47].

The dynamics of SM are influenced by the physical properties of topography and soil as well as temporal changes in atmospheric conditions. The impact of these parameters on the variability of SM has been studied in depth concerning topographic data [48]–[51], soil texture [51]–[53], and climate variables [54]–[56]. In general, the prediction of SM in local studies, e.g. station-based SM forecasting, does not require static parameters such as topography and soil texture since these data vary insignificantly. However, the variability of SM in time depends on climate data in both local, regional or global scale studies.

In the literature, researchers focused on minimizing the prediction uncertainties to estimate SM using *in-situ* measurements [57]–[61]. Including the meteorological parameters in estimating SM enhances the prediction accuracy significantly. The

[1]This chapter is based on : Celik, M. F., Isik, M. S., Yuzugullu, O., Fajraoui, N., & Erten, E. (2022). Soil Moisture Prediction from Remote Sensing Images Coupled with Climate, Soil Texture and Topography via Deep Learning. Remote Sensing, 14(21), 5584. https://doi.org/10.3390/rs14215584

study conducted in [58] predicted the SM values of five stations located in Shandong Province of China using varying depth measurements of SM together with the meteorological variables. A similar study was performed in [59], extending the spatial distribution of stations worldwide, to forecast the SM values. In this study, however, time series of each station trained and validated separately. Another study carried out by [60] used the SM values of globally distributed stations of International Soil Moisture Network (ISMN) coupled with climate, topography, and soil texture data to create a model for the daily prediction of SM in different depth layers. By spatially interpolating SM values of stations to form 0.25° grid cells, the trained model can predict SM in a quasi-global extend. Although the sensor measurements provide more reliable estimation of SM values, the dependency of model on SM sensors limits the use of model within specific regions where *in-situ* measurements exist. The lack of measurements in high latitudes resulted in poorer forecasts of SM values, specifically in arid regions.

Even though *in-situ* measurements play a crucial role in understanding SM, their spatial coverage and network-related problems make them limited in global studies. Recent developments in satellite-based remote sensing allowed continuous monitoring of the Earth's surface. In order to overcome the problems encountered SM predictions using *in-situ* measurements, satellite data from microwave remote sensing has been used excessively [62,63]. In this context, satellite images are the key to breaking free from the dependency of SM prediction from *in-situ* sensors. The data from the NASA-Soil Moisture Active Passive (SMAP) [64] and ESA-Soil Moisture and Ocean Salinity (SMOS) [65] missions are valuable asset for the global SM monitoring with their 2-3 days temporal resolution. In 2020, [66] expanded the near real-time SM predictions by integrating time series data from SMAP and SMOS missions using a statistical approach to overcome the inconsistencies between the different SM retrieval algorithms.

Although SMAP and SMOS SM products enable the monitoring of Earth's surface moisture in high temporal resolution, their applications are constrained due to their coarse spatial resolution. To overcome this limitation, researchers [67,68] used downscaling methods by merging higher resolution satellite images with lower

resolution SMAP/SMOS data to achieve improved spatial resolution SM predictions. Even though these downscaling efforts are applicable in predicting SM, the generated maps still have an insufficient spatial resolution ($\sim$5.6 km) for applications such as agricultural monitoring. In this regard, the launch of the Sentinel SAR satellites by ESA under the Copernicus Programme paved the way for accurate SM retrieval in smaller scale by acquiring higher spatial resolution microwave remote sensing images [69]–[73].

SM retrieval from remote sensing images has been improved by the state-of-art machine learning-based regression techniques owing to their ability to learn the relationship between predictors and SM from data [74]–[77]. An extensive review on the use of machine learning algorithms for predicting SM can be found in [78]. As computers have improved in performance, deep learning (DL) algorithms have become increasingly popular, as they can handle nonlinear and complex relationships between input and output [79].The SM forecasting studies that use remote sensing images exploited the ability of DL models to capture the spatial and temporal dynamics of SM at the expense of large datasets and high computation costs [45,80]–[85].

Among the different DL methods, artificial neural networks (ANN) have been carried out to estimation of SM from microwave remote sensing images integrated with some auxiliary data [86]. For example, while in [87] coupled S1 images with soil texture information, [84] used soil texture and soil temperature data to improve the prediction accuracy of SM retrieval. As an alternative to soil texture data, in [88] include climate and topography data to the ANN model. Further, in [82], the combination of soil texture, topography, and climate data was utilized to improve the ANN model's performance.

Recurrent Neural Network (RNN) is a DL technique that consider the sequential relationship between input and their effects on the output data. Therefore, such DL models are more appropriate when the sequence modeling tasks are needed, such as SM prediction. However, RNN suffers to learn inter-dependency between input and output when the sequence span gets longer [89]. In order to overcome the limitation of this DL technique, a special kind of RNN, long short-term memory (LSTM) is

proposed by [90]. With the LSTM, information from a sequence can be carried along the consecutive sequences, and the model can learn the relation between sequential data and output data.

The study conducted by [91] applied LSTM architecture for the first time in SM studies by using SMAP L3_SM_P product with climate and soil texture data to improve the design accuracy of SMAP SM data. In 2018, [92] presented a model for the long-term SM forecast on both surface and different depths over the continental US, aiming to exploit the SMAP data together with the land surface models. The model can predict long-term SM values in the same region using the SMAP SM time series data. In [93], the LSTM model trained with the same data classes used in [91] to nowcast the SM data, when SMAP L3_SM_P product become available. Another study [94] downscaled the SMAP SM data in ($\sim$1km) with the help of climate, soil texture and topography data by implementing LSTM.

This research aims to short-term SM prediction by combining high temporal resolution SMAP SM product and high spatial resolution S1 backscatter coefficients integrated with the auxiliary data to assist the agricultural activities in field-scale. In this context, we used the SM data of the ground stations from ISMN, distributed around the world, to train an LSTM model with two microwave radar data (SMAP and S1) together with soil texture, climate, and topographical data that are considered as the predictors of SM. The short-term forecast of SM on a field scale was successfully achieved by utilizing an approach dependent on microwave remote sensing satellite based observations. The model used in this study predict accurate SM values of the next day with high spatial resolution in regions with different geophysical properties and climate classes.

The manuscript is structured as follows: Section 4.2 explains the materials and methods; Section 3.4 describes the experimental research with data processing, model optimization, and our findings by focusing on the accuracy assessments of utilized methods; Section 3.5 presents the interpretation of the results and focuses on the effects of land cover, especially in vegetation presence, soil texture, and climate, on

SM estimation. We finalized the paper by highlighting the important outcomes of this study in Section 3.6.

3.2 Materials

In this research, we aim to predict SM by combining the satellite-based data (S1 and SMAP) with soil texture percentages (clay, silt, and sand), topography (elevation, slope, aspect, and hillshade), and climate (temperature, evapotranspiration, and precipitation). Using the features presented in Table 3.1, we modeled the SM in time using an LSTM framework. The statistics of these features were presented in Table 3.2.

3.2.1 International soil moisture network

ISMN is a data hosting facility developed and still maintained by several universities [95]–[97]. It is supported by the European Space Agency (ESA) Earth Observation program. The ISMN stations include soil texture properties and SM values in time, freely available at https://ismn.geo.tuwien.ac.at/. When we started the algorithm development, the total number of available stations was 1611 after 2017, when S1 data became available. The locations of the stations cover different climates and ecoregions. However, $\sim$70% of the available stations were located in the USA, see Figure 3.1.

Besides the station locations, in Figure 3.2, we present the ternary distribution of the soil data. Ternary distribution depicts the data in a 3D space, making it simpler to understand relations. Figure 3.2 shows that most soil samples are located in the loam class, followed by sandy loam, clay loam, and silty loam.

Along with the soil texture and SM data, the metadata of each station includes land cover based on the ESA CCI land cover product [98] and Köppen-Geiger climate classes [99]. It should be noted that these data were used only for evaluation of the model performance w.r.t. varying land cover and climate class of the stations, not for training the model.

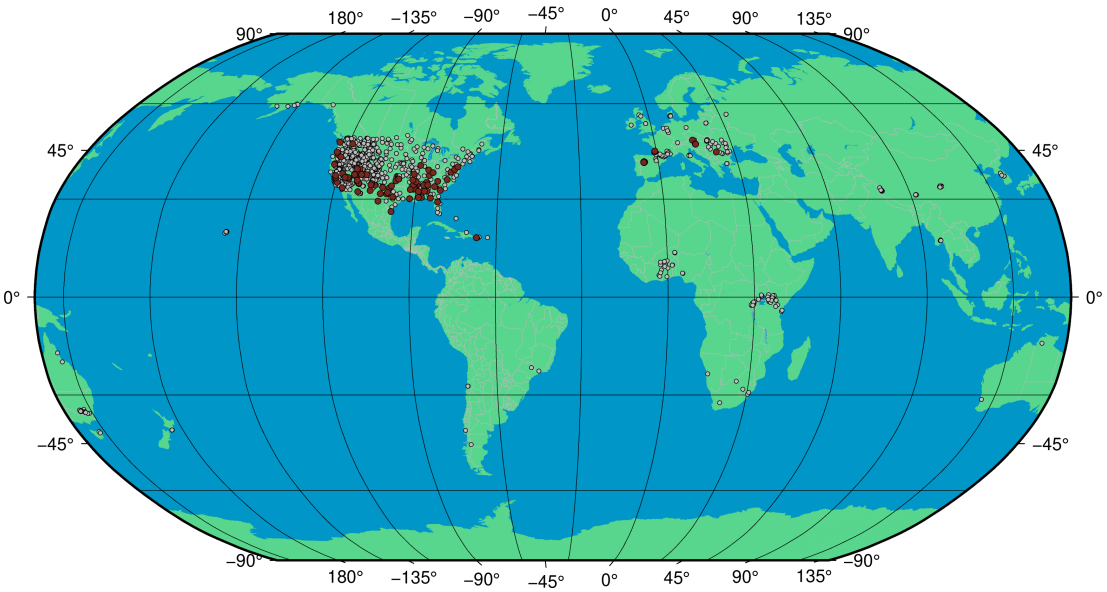


Figure 3.1 : The spatial distribution of ISMN sites. Red dots display the distribution of 103 stations with reliable data.

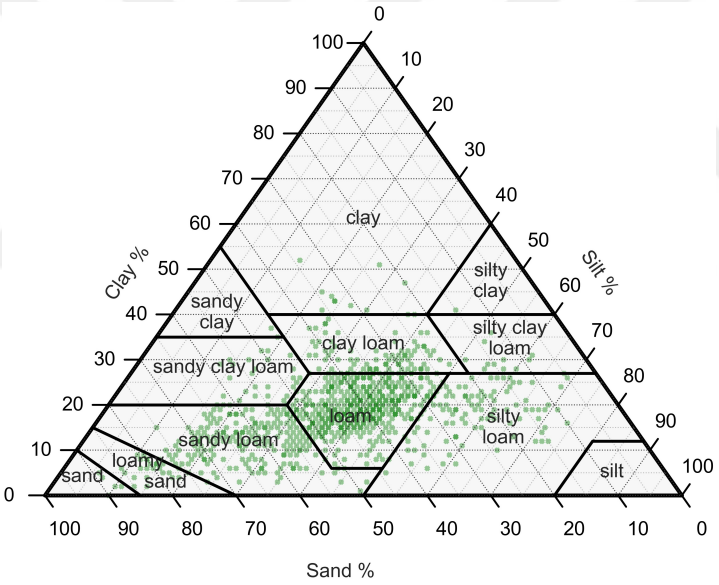


Figure 3.2 : Ternary plot of the soil class distribution of ISMN sites.

3.2.2 Satellite data

In this research, we accessed all satellite data via Google Earth Engine (GEE) Python Application Programming Interface (API) [100]. From the GEE, we downloaded the S1 data -one of the missions of ESA's Copernicus initiative- together with NASA's SMAP data on the location of the SM stations. Their ensured continuity for the future

and sensitivity to changes in vegetation and soil properties makes both satellites a viable option for SM monitoring [45,101]–[104].

3.2.2.1 Sentinel-1 (S1)

S1 is a Synthetic Aperture Radar (SAR) satellite mission with a C-band (5.6 cm) sensor. The advantage of S1 lies in its sensitivity to SM content [105]. There are two identical satellites in the S1 mission, S1a, and S1b. Each satellite has a temporal resolution of 12 days, resulting in an average of 6 days repeat cycle. Unfortunately, in December 2021, S1b failed data dissemination and became space junk. Since then, S1a has been providing data alone, and its temporal resolution depends on the area, with a minimum orbit repeat cycle of 6 days in Europe and 12 days in other areas. ESA is planning to launch S1c in the first half of 2023 to continue the dual satellite constellation.

This research used the Ground Range Detected (GRD) 10-meter spatial sampled data processed by ESA. The data we have selected has vertical transmission-vertical received (VV) and vertical transmission-horizontal received (VH) polarizations.

In this study, all S1 passes between 31 December 2017 and 01 January 2021 were included for each station of ISMN. In the data processing step, 50m*50m region of interest was defined around each station to calculate the mean value of S1 GRD backscatter signals. The mean backscatter signals were converted from logarithmic scale to linear scale. Additionally, VH/VV ratio was added as a feature to the dataset.

3.2.2.2 Sentinel-2 (S2)

S2 is a Multi-Spectral Instrument (MSI) satellite mission with spectral sensitivity to visible-near-infrared region of the electromagnetic spectrum. In this mission, like S1, there are two identical satellites (a and b). Both satellites have a temporal resolution of 12 days, also resulting in an average of 6 days repeat cycle.

In our research, we used the Level-2a surface reflectance product processed by ESA. The data has 13 bands ranging from 10 to 60-meter spatial resolution. We only used red and near-infrared bands to derive the vegetation indices. As in the case of S1, pixels

within the 50m*50m region of interest around the stations were extracted to calculate the mean NDVI values. However, this feature was only used to evaluate the model performance in the presence of vegetation and was not included in the feature set to train the model.

3.2.2.3 Soil moisture active passive (SMAP)

In 2015, NASA launched the SMAP satellite to monitor the SM content using L-band SAR (active) and radiometer (passive) instruments. SMAP has a temporal resolution of 2-3 days globally. In this research, we used Level-3 data of SMAP SM, which has 10 km spatial resolution [106].

3.2.2.4 Topography

The topography of the surface also influences the variation in the SM. With the GEE platform, topographic parameters such as elevation, slope, aspect, and hill-shade are obtained from the ALOS DSM Global 30m dataset [107].

Table 3.1 : Data used in this research provided with its descriptions, spatial and temporal resolutions.

Category	Feature Description	Spatial Res.	Temporal Res.
Climate Data[1]	$T(°C)$, $ET(mm)$ & $P(mm)$	1 to 5 km	Daily
Satellite Data[2] (S1)	VV, VH& VH/VV	10 m	6-12 days
Satellite Data (SMAP)	$Surface\ SM\ (mm)$& $Subsurface\ SM\ (mm)$	10 km	3 days
Soil Texture	$Sand$, $Clay$, $Silt\ (\%)$	Point-wise	Constant Values
Topographical Data[3]	$H(m)$, $S(°)$, $A(°)$, $HS(°)$	30 m	Constant Values
Soil Moisture Data	SM of top 5 cm (m^3/m^3)	Point-wise	15 mins

[1] T: temperature, ET: evapotranspiration, P: precipitation, [2] S1 backscatter coefficients in linear scale, [3] H: elevation, S: slope, A: aspect, HS: hillshade.

3.2.3 Climate data

As an integral part of the water cycle, the dynamics of SM are closely associated with climate data, such as precipitation, temperature, and evapotranspiration. In this research, we gathered the precipitation (P), air temperature (T), and evapotranspiration (ET) data on the location of the SM stations using the Meteomatics API [108]. The available meteorological data have a spatial resolution ranging from 1km to 5km.

Under the assumption of lower spatial variability, we used the reported data without changing the processing pipeline. The usage of the API was made possible within the service provided to AgriCircle AG by Meteomatics.

3.2.4 Data preprocessing

For SM modeling, we created a dataset that combines static and dynamic features, as previously shown in Table 3.1. The static features are soil texture and topography; the dynamic features are climate and satellite-derived time-series data. In addition, we added time variable as a dynamic feature. Since the LSTM framework requires time-series data, we repeated the static features as the sequence length before feeding it to the LSTM framework.

For dynamic features, we prepared a three-year data set that includes *in-situ* observations acquired from ISMN stations from 31 December 2017 to 01 January 2021. In this dataset, we applied data cleaning to reduce the data originated uncertainty and eliminate the inconsistency within the measurements. Data cleaning involves two-step elimination criteria. The first criterion is related to the record length. The record length condition requires that those stations be discarded if more than 10% of the measurements were missing in any station. The second criterion is developed to ensure sequential dependence in the observations. The SM stations with more than 60 consecutive days of missing measurement are also excluded from the analysis since a solution like interpolation was unrealistic considering the complex nature of the problem. According to these criteria, we found 103 stations, shown by red dots in Figure 3.1, out of 1611 with time series of SM measurements suitable for the analysis. Since dynamic features are gathered from various sources with different temporal resolutions, we upsampled all data into daily sampling using the linear interpolation method for temporal matching. The ground measurements are re-sampled into daily SM values to ensure the matching temporal resolution.

For the training of the LSTM model, we formed five different scenarios to determine the contribution of feature groups. As previously shown in Table 3.1, in SM monitoring, climate data, soil texture, and topographical data are the main drivers of SM. Beginning with the climate data (Case-I), we consecutively included soil texture

(Case-II), topographical data (Case-III), and satellite data (Case-IV and Case-V) and listed them below.

- Case-I : Climate data
- Case-II : Climate data, soil texture
- Case-III : Climate data, soil texture, topographical data
- Case-IV : Climate data, soil texture, topographical data, satellite data (SMAP)
- Case-V : Climate data, soil texture, topographical data, satellite data (SMAP, S1)

In each case, time variables (sine and cosine of time) are kept within the features set since they are independent variables that represent the positional encoding of input features in a time series.

Table 3.2 : The statistics of features used in the study.

Feature	**Mean**	**Std**	**Feature**	**Mean**	**Std**
Temperature (T)	8.81	11.03	Sand	42.81	13.91
Evapotranspiration (ET)	2.80	1.96	Clay	18.77	6.90
Precipitation (P)	2.64	11.01	Silt	38.42	10.87
VV	0.019	0.019	Elevation (H)	1400.48	1150.57
VH	0.088	0.076	Slope (S)	7.55	7.16
VH/VV	0.229	0.281	Aspect (A)	162.99	104.83
SMAP SM (Surface)	14.70	8.62	Hillshade (HS)	180.10	23.09
SMAP SM (Subsurface)	52.56	37.97	Soil Moisture (SM)	0.18	0.12

3.3 Methods

We employed the satellite data, soil texture, climate, and topography features mentioned above to forecast the SM using the following process chart shown in Fig. 3.3. The process starts with the first row and ends with the accuracy assessment and prediction of SM.

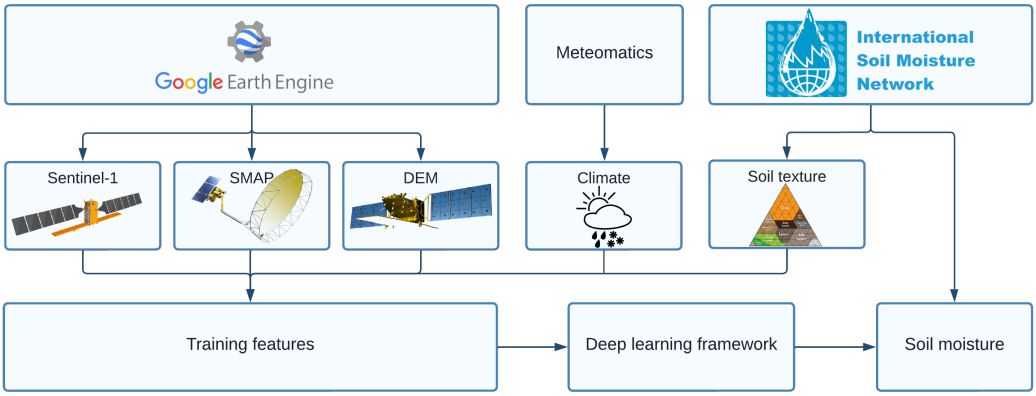


Figure 3.3 : The overall process chart of the study, starting from data sources and ending with the final-user output.

3.3.1 Long short-term memory

As a descendent of RNN, [90] proposed an approach called Long Short-Term Memory (LSTM) to overcome the vanishing gradient problem in RNN. In LSTM, the ordinary unit cell repeats the input-output sequence; in RNN, this is replaced by a memory cell. LSTM contains three gates: the input gate i_t, forget gate f_t, and output gate o_t. Besides these gates, there are two different parts: cell state c_t, which keeps information from previous states and transfers it to the next, and the hidden state h_t that the output of the LSTM cell. The equation of input gate, forget gate, and output gate is defined as;

$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{3.1}$$

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{3.2}$$

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{3.3}$$

where w_i, w_f, and w_o are weight matrix, x_t is input, h_{t-1} is the hidden state from previous time step, b_i, b_f and b_o are bias vector and σ is the sigmoid activation function for the gates. The activation functions introduce non-linearity by transforming inputs to targeted outputs with a nonlinear regression procedure, making the model capable of learning and performing more complex tasks. After the calculation of gates, the cell state and hidden state can be defined as;

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(w_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{3.4}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{3.5}$$

where w_c is weight matrix, c_{t-1} is cell state from previous time step, b_c is bias vector, *tanh* is hyperbolic tangent activation function and $\odot$ is element wise multiplication. The size of the weight matrix is determined according to the unit size and hidden layer size of the LSTM model, feature vector dimension, and feature sequence length. It should be noted here the weight matrix of LSTM does not change through timesteps. For detailed information please refer to [109].

3.3.2 Accuracy assessment

Four accuracy metrics, namely, coefficient of determination (R^2), root mean square error *(RMSE)*, unbiased root mean square error *(ubRMSE)*, and mean absolute error *(MAE)* were used to evaluate the performance of the implemented model for the SM prediction.

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \\ MAE &= \frac{\sum_{i=1}^{N}\left|y_i - \widehat{y}\right|}{N} \qquad ubRMSE = \sqrt{(RMSE)^2 - (\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right))^2} \end{aligned} \tag{3.6}$$

In the above equations, y_i, $\hat{y}_i$, and $\bar{y}_i$ indicates actual SM, predicted SM and mean value of the actual SM, at ith time step respectively. Out of these four metrics, we use R^2, *RMSE*, and *ubRMSE* to evaluate the performance and *MAE* for station-based assessments of the trained model.

3.3.3 Implementation of the LSTM framework

The SM value at time t (Y_t) was predicted by using n number of input features with previous w sequential days (window size) as $[X_{t-1}^n \ldots X_{t-w}^n]$. After preparing the dataset, we divided it temporally into 60% for training, 10% for validation, and 30% for testing purposes. The temporal split corresponds to 658 days used to train the model

starting from 31 December 2017 until 20 October 2019, 109 days used to validate the model training between 21 October 2019 and 06 February 2020, and 330 days used to evaluate the trained model from 07 February 2020 until 01 January 2021. While the LSTM model was built with training data, the hyperparameter tuning was carried out by using validation dataset. After the optimum hyperpamater set was determined, independent evaluation of the model was conducted based on testing data.

Before starting the training, we normalized all the input features via *MinMaxScaler* function of *sklearn* python package to ensure numerical stability. For the normalization, we followed different strategies for static and dynamic features. By their nature, the static features have global minimum and maximum values; therefore, we normalized them together. On the other hand, dynamic features have local variations that change each station's minimum and maximum values, leading to a station-based normalization.

One of the primary flexibility of using time series data is using a varying length of past data to make future predictions. In such a structure, the number of previous time steps is called the window size. The window size parameter must be selected carefully since it impacts forecast accuracy. For its determination in the SM forecast, we reformed the original dataset according to different window sizes: last one day, five days, ten days, and thirty days.

The LSTM networks were created using *TensorFlow* back-end with GPU processing integration in *conda* environment. We used the *gridSearchCV* function of *sklearn* python library, to determine the LSTM model's hyperparameters. Besides, in LSTM architecture, all models started with an LSTM layer, followed by a one-dimensional dense layer as an output.

3.4 Results

The results of the SM prediction framework were presented in this section, starting with data preparation followed by model training, model parameter optimization, and finalized with the assessment of feature effects.

3.4.1 Model parameter optimization

The grid search algorithm was applied with using various hidden layers and unit sizes, learning rates, loss functions, and optimization functions for hyperparameter optimization. The number of hidden layers for LSTM was tested by gradually increasing from a single layer to three stacked layers. The unit size of these stacked layers was tested for 32, 64, and 128. The tested learning rates were 10^{-2}, 10^{-3}, and 10^{-4}. For optimization function, we tested Adam, Adamax, and SGD [110]. For epoch number, the test was for values between 1000 and 1500 with 100 steps. Lastly, the dropout rate was between 0 and 0.5 with 0.05 increments.

The performances of the trained models with setups having different window sizes are presented in Table 3.3. We can see that the window size of 5 days is performing better than other window sizes, with the overall *MAE* reduced to ~ 0.03 for both training and testing. Out of these four different window sizes, the 1-day window size showed the worst prediction results with R^2 values of ~ 0.70 for both training and testing. Following the window size of 5 days, 10, and 30 days gave comparable results.

Table 3.3 : Accuracy of LSTM models with different window size.

Window	Train				Test			
Size	R^2	RMSE	ubRMSE	MAE	R^2	RMSE	ubRMSE	MAE
1	0.701	0.069	0.069	0.053	0.695	0.071	0.071	0.053
5	**0.922**	**0.035**	**0.035**	**0.026**	**0.871**	**0.046**	**0.045**	**0.033**
10	0.922	0.035	0.044	0.026	0.859	0.048	0.048	0.035
30	0.900	0.040	0.040	0.029	0.837	0.052	0.048	0.038

Focusing on the window size of the last five days, which performed better than the other tested cases, we found that LSTM with two hidden layers and 32 unit sizes followed by a one-dimensional dense layer having a learning rate of 10^{-3}, epoch number of 1000, the dropout rate of 0.25 and Adamax as the activation function gave the best accuracy for SM prediction. The summary of the grid search is given in Table 3.4.

Table 3.4 : Hyperparameter ranges of LSTM model and selected values for the last 5 days window size.

Hyperparameters	Tested	Selected
Hidden Layer	1,2,3	2
Unit Size	32, 64, 128	32
Learning Rate	0.01, 0.001, 0.0001	0.001
Activation Function	Adam, Adamax, SGD	Adamax
Epoch Number	1000 - 1500	1000
Dropout Rate	0 - 0.5	0.25

3.4.2 Effect of the different features on the model performance

After the optimum window size and hyperparameters were assessed, we investigated the effect of a different group of features on the model's prediction capability by designing five different cases. Table 3.5 summarizes the statistics of these cases for their corresponding feature combinations where the model hyperparameters are based on the best performing LSTM model with a window size of 5 days (see Table 3.4). We found that the optimum solution for SM prediction was achieved when all feature groups were combined, i.e., Case-V, for training the LSTM model.

Table 3.5 : Accuracy analysis of LSTM with different features set.

Case No	Train				Test			
	R^2	RMSE	ubRMSE	MAE	R^2	RMSE	ubRMSE	MAE
Case-I	0.366	0.101	0.101	0.082	0.337	0.105	0.104	0.085
Case-II	0.663	0.074	0.074	0.057	0.651	0.076	0.076	0.058
Case-III	0.875	0.045	0.045	0.033	0.843	0.051	0.051	0.037
Case-IV	0.908	0.038	0.038	0.028	0.860	0.048	0.046	0.034
Case-V	**0.922**	**0.035**	**0.035**	**0.026**	**0.871**	**0.046**	**0.045**	**0.033**

3.4.3 Overview of the model training

Figure 3.4 presents the training progress of the best performing LSTM model, whose optimum hyperparameters are given in Table 3.4. The figure shows the change in the loss value, R^2, and *RMSE* w.r.t. epoch as the model continues its training with a constant learning rate of 10^{-3}. The loss value, R^2, and *RMSE* for training and

validation datasets converge around epoch number 1000, and the model tends to over-fit beyond 1000 epochs.

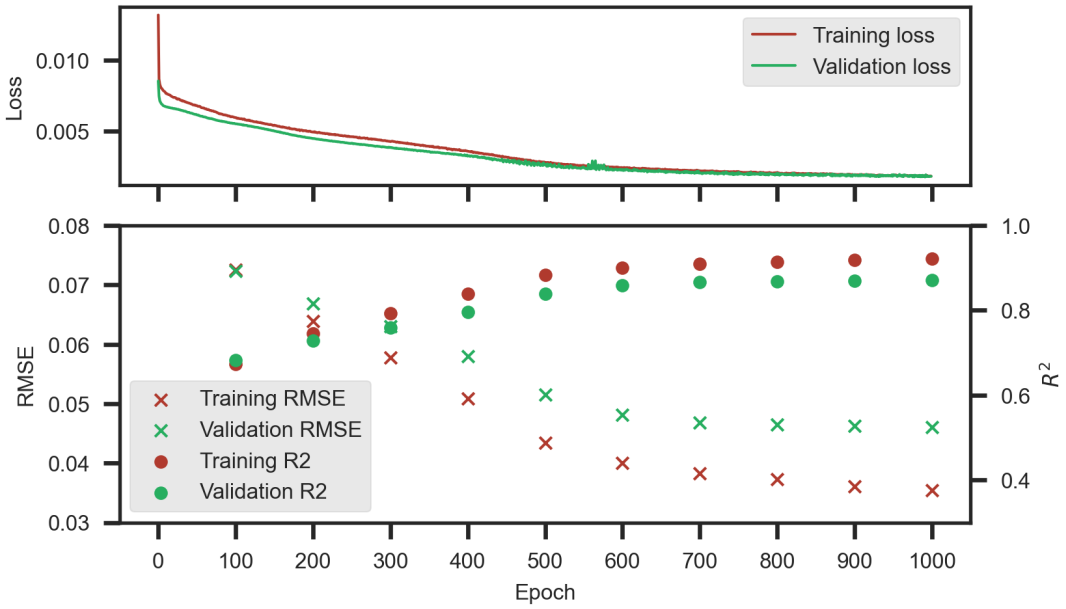


Figure 3.4 : Accuracy of the best performing LSTM model according to epoch. The upper figure shows the training progress of the model w.r.t. loss value per epoch, and the lower figure shows the change in accuracy w.r.t. R^2 and *RMSE*.

Figure 3.5 shows the outcomes of the training (left-side) and testing (right-side) SM predictions for all stations. The scatter plots between measured and estimated values for the training and testing datasets show a similar pattern when compared. The main population of the points is along the 1-1 line. The model can make good predictions with *MAE* of less than 0.035. In the second row, violin plots show the measurement and prediction distributions. The left side of the violin corresponds to actual values, while the right side stands for the predictions. In an ideal case, we should see a mirror-like shape, which is also the case for our predictions with small differences due to the error previously mentioned in the scatter plots.

3.5 Discussion

The LSTM-based SM forecast model relies on satellite-driven data, soil texture, topography, and climate. Therefore, as the predictions are conducted for different conditions, we investigated the prediction performances for land cover classes,

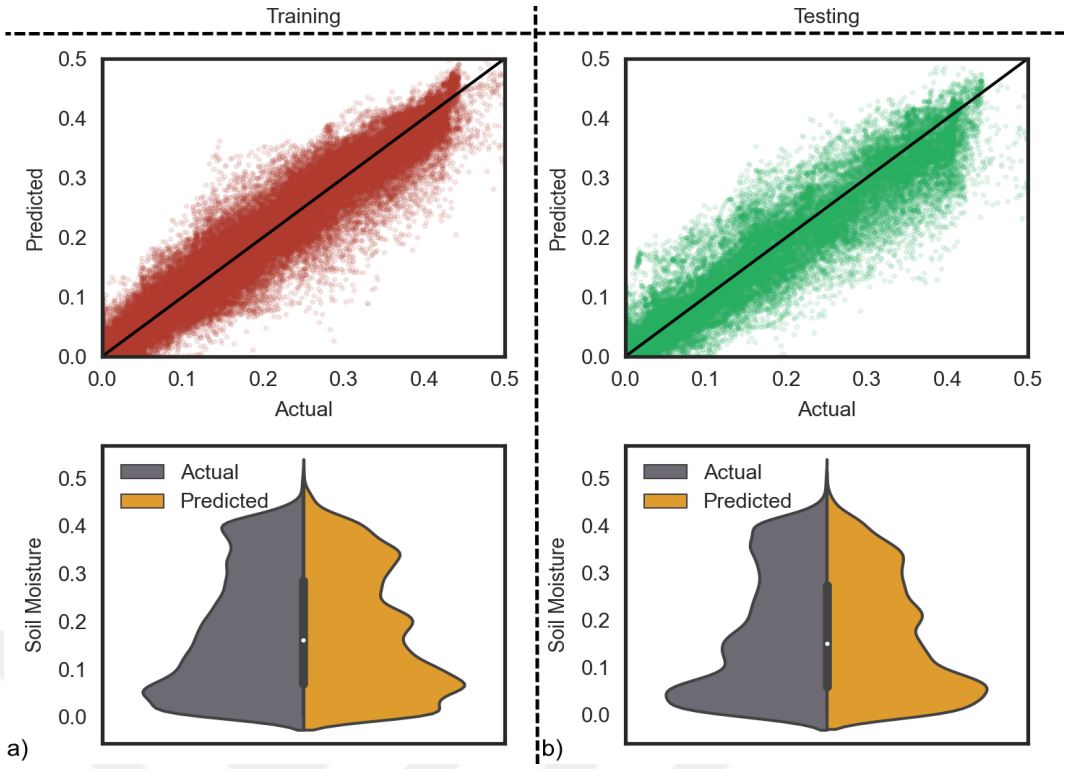


Figure 3.5 : The scatter plot (top left and right) and distribution graph (bottom left and right) of (a) training and (b) testing data of windows size 5.

biomass variations based on the NDVI calculated from the Sentinel-2 satellite, climate classes, and soil texture.

3.5.1 Relationship between model performance and land cover

The physical characteristics of the land cover affect the prediction accuracy of the developed LSTM model. This effect originates from the physical heterogeneity of the observed area.

In the ISMN, every station is provided with its land cover type. The corresponding land covers are based on the ESA CCI land cover product [98]. In a total of 103 stations, 34 croplands, 20 grasslands, 18 shrublands, 23 trees/forest, and 6 mosaics (mixture of trees, shrubs, herbaceous, and cropland), and two urban sites exist. However, we did not investigate the urban sites due to the insufficient number of samples.

Figure 3.6 presents the model's prediction capability for different land covers. The smallest *MAE* (~ 0.02) was achieved for shrubland class. The model shows similar performance for cropland, grassland, and tree covers with a mean *MAE* around ~ 0.03.

However, the variance of *MAE* for the cropland cover is higher than the others. The worst *MAE* (~ 0.05) is obtained for the mosaic cover due to the complexity of the surface. This can be explained by the scattering mechanism of SAR imagery in the presence of vegetation and forest. Since the shrubland land cover class is sparsely vegetated area, radar signals can interact with the soil more than vegetation or forest canopy.

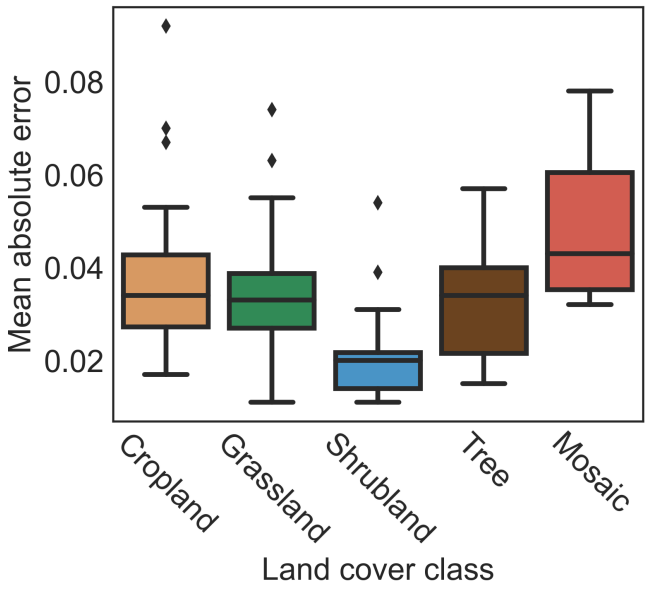


Figure 3.6 : Overall *MAE* for land cover classes.

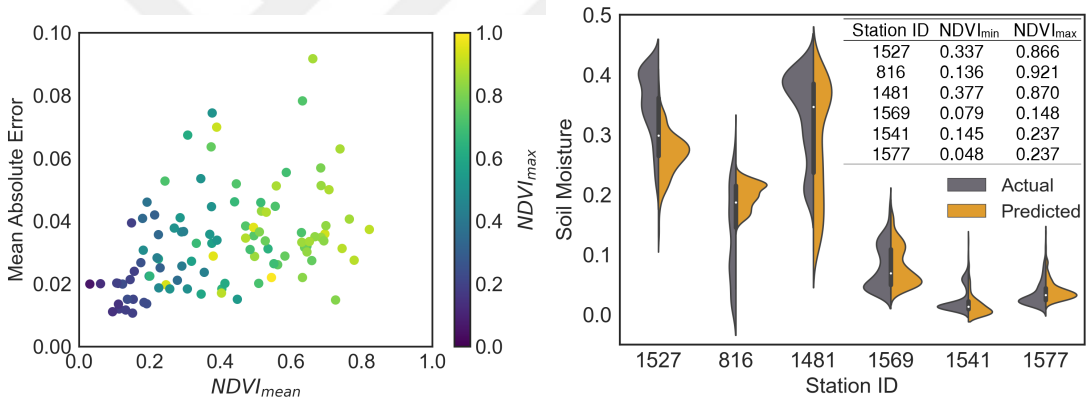
3.5.2 Relationship between model performance and NDVI

The presence of biomass over soil may affect the model's prediction capability since the satellite data also carries information regarding the vegetation. To see the effect of the biomass, we calculated the NDVI from the S2 surface reflectance image during the testing periods and compared it with the MAE values of the model for the prediction dates.

Figure 3.7a visualizes the distribution of *MAE* values for all available stations together with the $NDVI_{mean}$ and $NDVI_{max}$ values. The figure shows the correlation between the mean $NDVI_{mean}$ and *MAE* values. MAE values tend to increase with increasing $NDVI_{mean}$ values.

The violin plot given in figure 3.7b shows the distribution of the actual vs. predicted SM values at stations whose *MAE* values are lower (Station ID: 1569, 1541, 1577) with low soil moisture and higher (Station ID: 1527, 816, 1481) with high soil moisture. Here, we focused on finding out the origins of the variations in *MAE* values among these stations. For this purpose, the variation of the NDVI values were used. This analysis showed that the NDVI variation is one of the reasons for the deterioration of the SM prediction.

The backscattered signals obtained from SAR data were strongly affected by high biomass due to the interaction between electromagnetic radiation, plant, and soil. Therefore, these findings show that the model's estimation performance is prone to uncertainties from the existing biomass. Similar findings also exist in the previous studies [75,111]–[113]. These studies found that the SM content in bare or low-density vegetation areas is more predictable than in high-density vegetation areas.



(a) (b)

Figure 3.7 : Model performance w.r.t. NDVI variation, (a) scatter plot shows the distribution of *MAE* vs. *NDVI* relationship for each station, (b) Violin plots representing the statistical distribution of actual and predicted temporal SM data at the ISMN stations with their minimum and maximum NDVI values.

Another investigation that we conducted on the impact of NDVI variation was using station-based time series. For this purpose, we focused on some stations that show a variation in NDVI over the years. We see that the growth cycle of NDVI values before seeding and after harvest is lower than crops' vegetative and reproductive phases. We believe that the prediction capability of the model though out the growth cycle is an

important detail that needs to be investigated. Hence we prepared the Figure 3.8a to show the model's performance in time. According to Figure 3.8a model's performance on the SM forecasting dropped approximately between May 2020 to October 2020 due to very low SM values. During this period, we can see an increase in the NDVI values from ~ 0.2 to ~ 0.9. We observed a similar situation in the other stations as well. In the time series of stations 827 and 1572, given in Figure 3.8b & 3.8c, the station has higher NDVI values from June to the end of December and from mid of April to the beginning of November, respectively. These three stations and the others with similar behavior have *MAE* values less than 0.075.

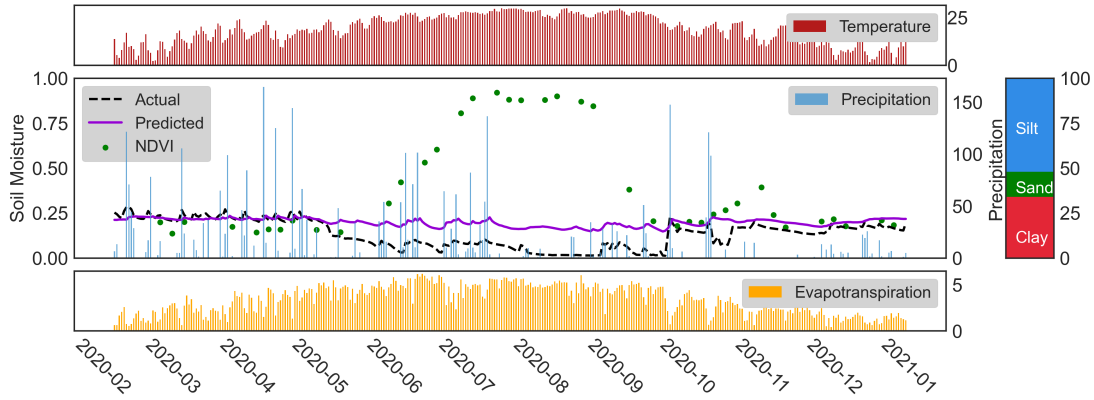
3.5.3 Relationship between model performance and soil texture

The variation in the soil texture is a driving factor for the spatial and temporal changes in the SM. Soils with high clay or silt fraction are associated with high water holding capacity, resulting in a generally higher SM value. On the other hand, such soils lose their moisture slower than the others. From an agricultural point of view, clay soils have the highest soil moisture content in general; however, silty soils are more favorable for plants.

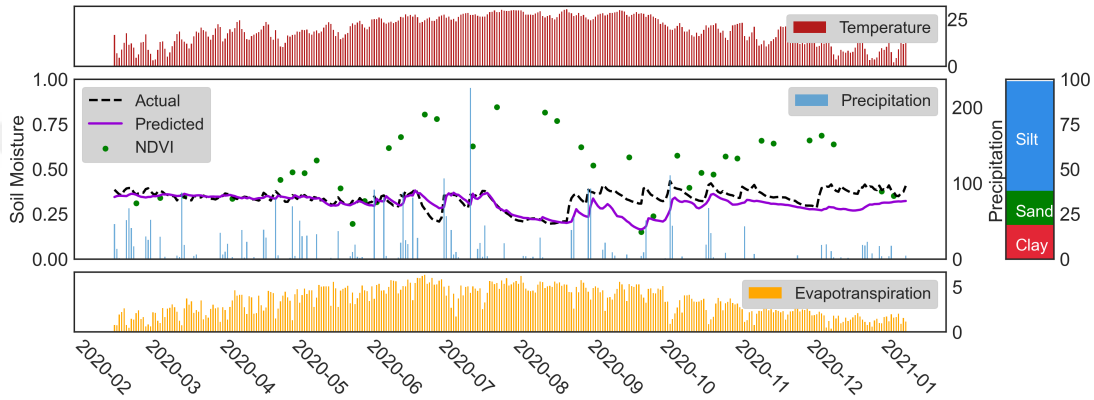
We provide a ternary plot in Figure 3.9 to show the *MAE* values of stations, which are scattered based on their soil texture contents. In the same figure, we also included each station's $NDVI_{mean}$ values in a color map. The combination of soil texture and $NDVI_{mean}$ allows us to observe the relationship between the amount of silt and clay in the soil and vegetation activity.

The size of each circle, representing a station, is proportional to its *MAE* value. We observe that the smaller circles generally accumulate in areas where the sand fraction is high. Among all the stations, 61% have sandy soil with an average *MAE* of 0.03, and 38% of them are silty soils with 0.04 average *MAE*.

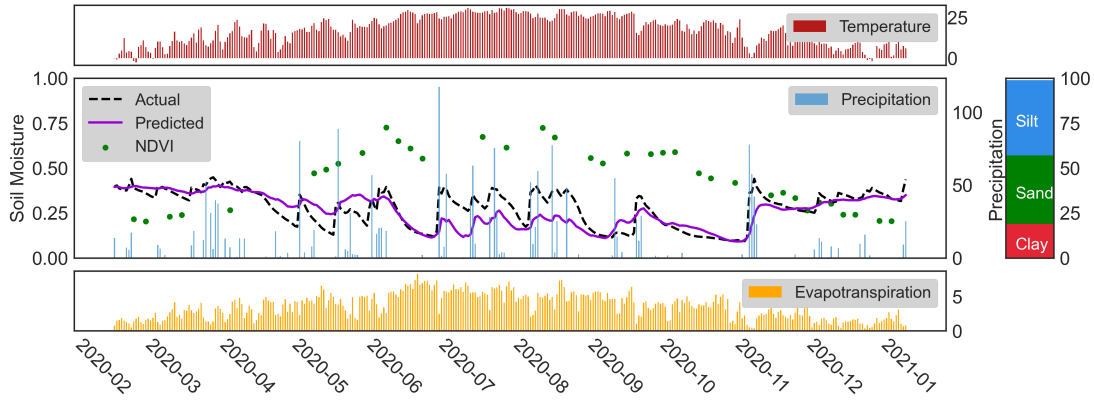
As we focus on particular stations for an in-depth investigation, it was observed that the silt content of the stations, having cropland cover, given in Figure 3.8 are 52%, 61% and 42% for stations 816, 827 and 1572, respectfully. In the corresponding stations,



(a) Station ID: 816



(b) Station ID: 827



(c) Station ID: 1572

Figure 3.8 : Time series of SM predictions during the testing period for stations 816, 827, and 1572.

we have similar findings that justify the performance of the model w.r.t. the change in the NDVI values.

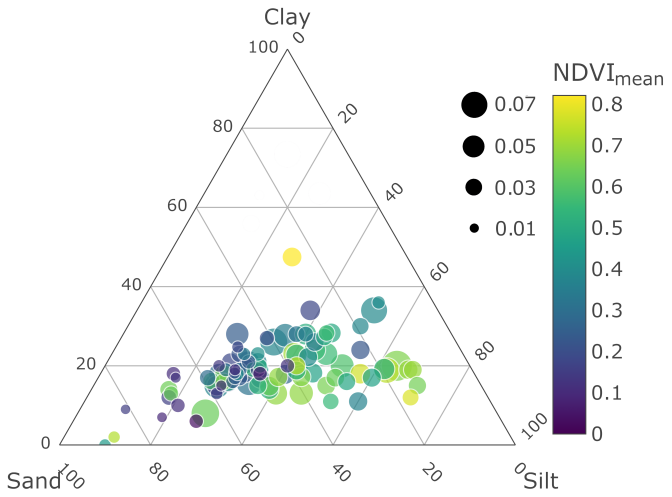
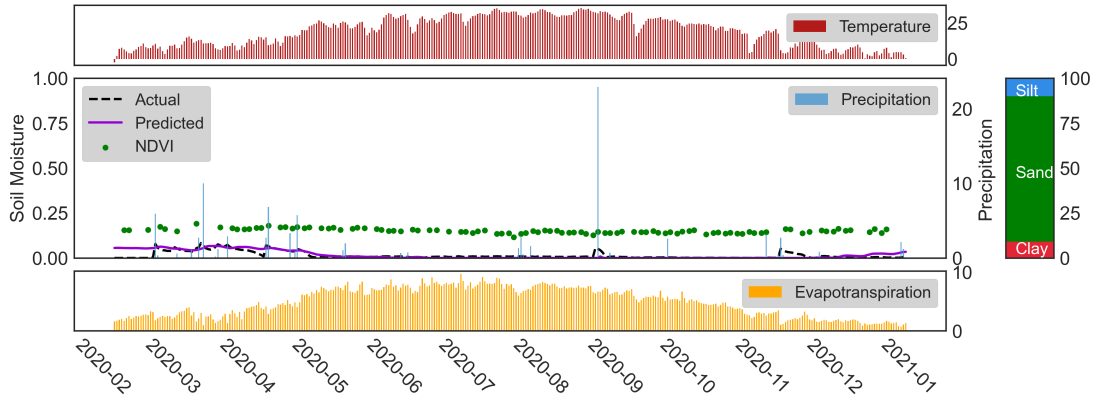


Figure 3.9 : Soil texture ternary plot w.r.t. MAE of each station. The circles are scaled based on their MAE value and are colored based on $NDVI_{mean}$.

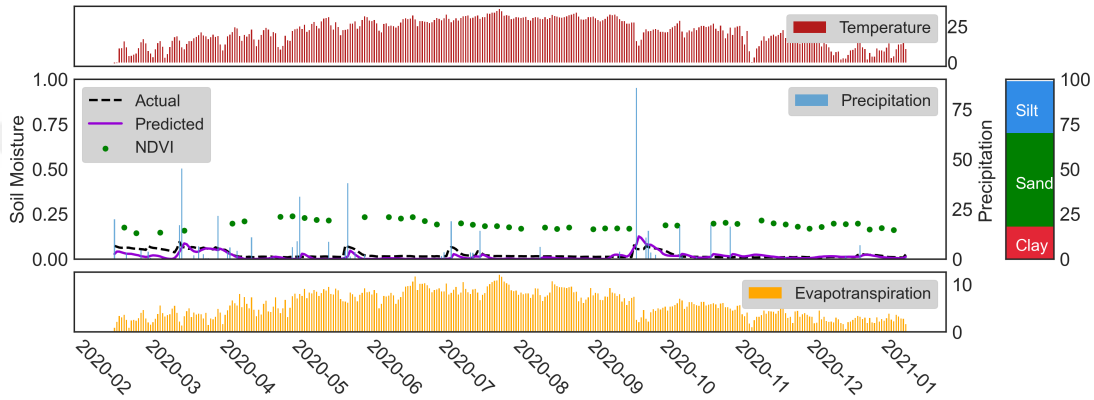
Besides silt and clay-dominated soils, the soil types in which the sand proportion is higher generally have a lower trend in SM values since the sandy soil has low water holding capacity. This property makes them less suitable for agricultural applications. In order to investigate the sand effect, we present the time series of SM predictions at stations 815, 1541, and 1569 in Figure 3.10. The typical features of these stations are the high percentage of sand fraction in soil content (81%, 52%, and 52% for stations 815, 1541, and 1569) and lower NDVI values along the time series. The mean NDVI value for these stations is 0.15, 0.19, and 0.11, respectively. Unlike the findings from Figure 3.8, we saw that in Figure 3.10a, the higher sand fraction leads to lower and less fluctuated SM values. Thus, the highest accuracy was obtained at stations with sandy soils having low NDVI values.

3.5.4 Relationship between model performance and climate classes

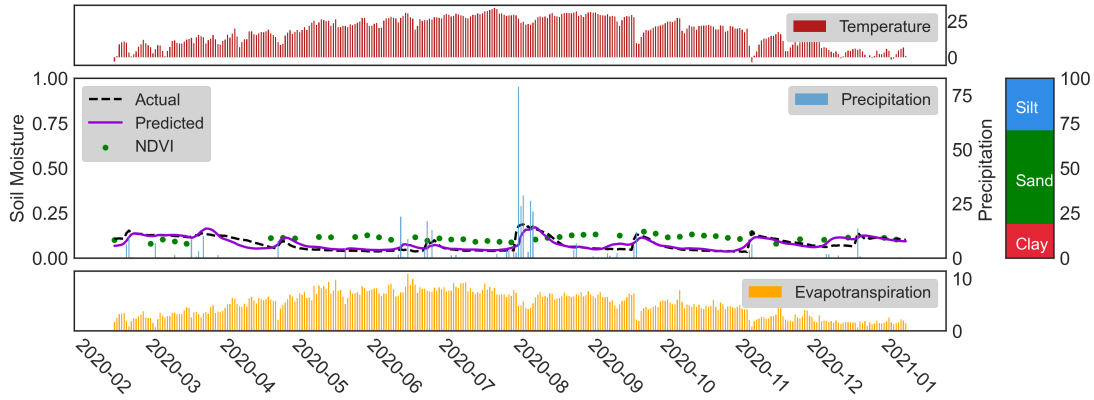
Lastly, we investigated the effect of climate classes. To this aim, we used [99], which defines four classes in total: tropical (A), dry (B), temperate (C), and continental (D). Our selected stations are distributed as 23% in class B and 75% in class C. The remaining 2% belongs to classes A and D, with one station for each.



(a) Station ID: 815



(b) Station ID: 1541



(c) Station ID: 1569

Figure 3.10 : Time series of SM predictions during the testing period for stations 815, 1541, and 1569.

In Figure 3.11, we present the model's prediction performance under different climate conditions as a boxplot. The stations in class B shows lower *MAE* values compared to those in class C (see Figure 3.11a). Considering the climate class properties, the

rapid changes in the moisture affect the dielectric properties of the target [72,114]; at the same time, precipitation is a significant factor that negatively impacts the SM prediction due to the change in the interaction between SAR signals and land surface.

We obtained better soil moisture predictions in arid climates (Bw) than those in semi-arid climates (Bs) regions due to less precipitation and more evapotranspiration. We also observed a similar behavior between no dry season climate (Cf) and dry summer (Cs) temperate climate classes (see Figure 3.11b). While no dry season climate, as inferred by its name, has a high precipitation rate compared to a dry summer climate, which makes the stations located in this climate region challenging in SM prediction.

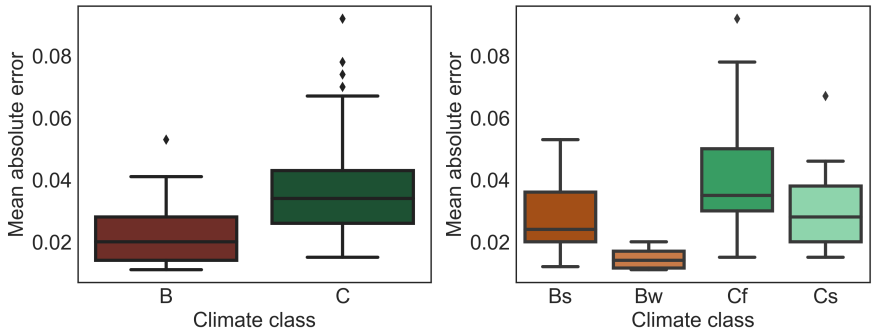


Figure 3.11 : Overall mean absolute error for first order (a) and second order (b) Köppen-Geiger climate classes [99].

3.6 Conclusions

In this study, we investigated the short-term SM prediction based on satellite-derived data with LSTM. For this purpose, the static and dynamic features were combined to create sequential input data and used *in-situ* SM measurements of 103 stations from ISMN as an output to train an LSTM model. Our approach uses soil texture and topographical data as static features and satellite (S1 and SMAP) and climate data as dynamic features. As SM monitoring is crucial for water resource management, we employed the SAR data due to its lower sensitivity to atmospheric conditions than optical data. To optimize the LSTM models' hyperparameters, we used *gridSearchCV* algorithm. After the optimization, the overall testing accuracy of the model was

calculated as $R^2 = 0.87$, $RMSE = 0.046$, and $MAE = 0.033$. The values obtained from different stations are summarized in Appendix A, including the station ID, network and station name, soil texture, NDVI mean and max values, climate, land cover classes, and the corresponding *MAE* values.

During our investigations, it was observed that the model's prediction performance is affected by the soil texture, vegetation status, and climate conditions. Variations in soil texture change the soil water holding capacity. In the case where the amount of sand was dominant, the SM values were easier to model than in the case of silt and clay dominance due to the low SM values and fewer fluctuations in sandy soils. We also observed that vegetation affects the interaction between the SAR signal and the soil. Thus, the model's prediction ability was lowered in vegetated areas with high NDVI values. Moreover, the model can predict better under dry climate conditions, such as arid and semi-arid climates in relatively low precipitation.

This study used satellite-based products to create a model to forecast SM values. For operational purposes, we know that obtaining soil texture data on the pixel level is challenging. However, we can overcome this by conducting an intensive sampling campaign for soil texture, or existing models can be used [115], which employs S1 and S2 multi-temporal data.

In the future, we plan to combine the LSTM model with the attention mechanism to study the contribution of each variable to SM prediction. The LSTM model combined with the attention mechanism can determine the importance of each feature and its temporal relationship with SM phenomena. Thus, we can increase the accuracy of the model predictions and explain the physical behavior of the black-box model.

4. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR COTTON YIELD PREDICTION WITH MULTISOURCE DATA[1]

4.1 Introduction

As a non-food crop, cotton is vital in supporting the textile industry, which has great power in the global economy. The sustainability of cotton production is of great importance, both from an economic standpoint and for preserving the ecosystem. Like other crops, cotton production is under threat due to climate change, extreme meteorological events worldwide, the amount of fertilizers and pesticides used, and water consumption, which negatively affects the soil. Consequently, cotton yield prediction needs to be significantly considered for crop production, land use decisions, and the management of economic impacts.

Satellite remote sensing images effectively observe crop conditions and growth cycles with derived biophysical parameters [22,113,116]–[118]. Biophysical parameters can facilitate the estimation of crop biomass and the monitoring of crop health, eventually aiding in predicting potential crop yield. The fluctuation in potential crop yield can stem from changes in climatic factors and soil moisture influenced by these factors. During the growth cycle of the crops, the variation in temperature, precipitation, soil moisture, etc., can cause physical damage and crop stress that eventually risks crop health and its development [119,120]. Aside from the crop-originated and climatic parameters, the variation in soil properties related to agricultural productivity determines the water-holding capacity, air circulation into the soil, and the relationship between crop and soil [121]. While meteorological parameters directly affect the crop growth cycle, which is a leading indicator of the yield to be obtained, soil properties

[1]This chapter is based on : Celik, M. F., Isik, M. S., Taskin, G., Erten, E. & Camps-Valls, G. (2023). Explainable artificial intelligence for cotton yield prediction with multisource data. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1007/s12145-022-00843-2

also have an essential effect on carrying out agricultural practices like irrigation and fertilization [122,123].

Making efficient agricultural decisions based on understanding the spatial and temporal variations of crop yield and their relation to changes in climatic and pedological conditions remains challenging. Developing a reliable yield estimation model that can assist farmers in agricultural planning requires an explicit interpretation of the functional relationship between environmental features and crop yield. In the last few decades, researchers have carried out many studies to predict yields with the combination of remote sensing images, meteorological data, and soil properties for various crops using remote sensing satellite images and artificial intelligence [124]–[126]. These studies attempted to increase the accuracy of yield prediction by exploiting the abilities of state-of-the-art shallow machine learning (ML) and deep learning (DL) approaches to solving complex dynamic problems. These models provide highly accurate results with their complex modeling capabilities. Yet, they are generally black boxes, which means it is challenging to understand how the predictors affect the model's behavior without the help of post-hoc methods that enable the explainability [127]–[130]. The trustworthiness of post-hoc methods in explaining the decisions of a black-box model, whether locally or globally, might be limited since they do not rely on the black-box model used. Furthermore, the reliability of the explanation provided by post-hoc methods varies greatly depending on the effectiveness of the particular post-hoc method.

Glass-box methods are preferable to black-box methods because they provide transparency, reliability, and ease of interpretation. While they may not be as accurate as black-box methods, they avoid the need for post-hoc explanations by allowing users to understand the model's inner workings. A glass-box method, known as explainable boosting machines (EBM) [131], has been proposed recently and has already been successfully applied to a range of phenomena, such as slope failure detection [132], deformation monitoring for concrete structures [133], and biomedical signal processing [134]. This study utilizes the EBM for the first time in agricultural studies. This approach provides competitive accuracy in predicting yields and allows a better understanding of the model's inner dynamics. We integrated remote sensing

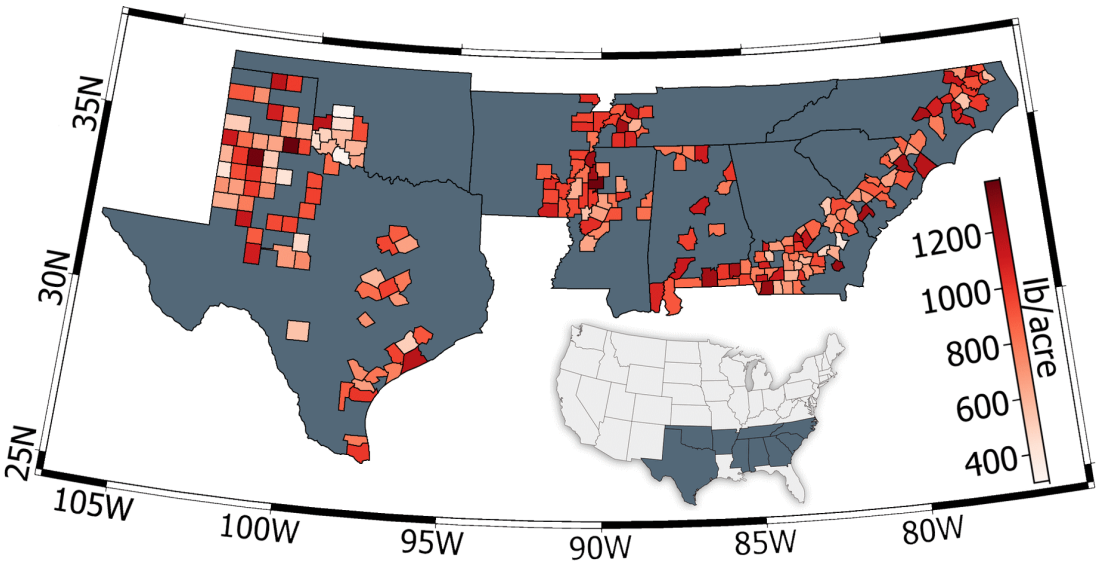


Figure 4.1 : Illustration of the study area along with the counties. The color bar corresponds to average yield records between 2017-2021.

satellite images, climatic factors, and soil properties to predict end-of-season and within-season cotton yields and to identify the key factors in the growth cycle affecting the obtained yield. The results of our study provide valuable insights into agricultural modeling.

4.2 Materials and Methods

4.2.1 Materials

This study covered nine states (Alabama, Arkansas, Georgia, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, and Texas) that account for 95% of yearly cotton production in the Continental United States (CONUS), the third-largest cotton producer in the world according to the United States Department of Agriculture (USDA). The USDA performs a thorough investigation of crop production in the CONUS annually and releases a high-resolution crop classification map [135]. The cotton classification map of these nine states and corresponding annual yield records were collected. In Figure 4.1, the counties included in this study are presented with their averaged cotton yield between 2017-2021.

The data gathered from multiple sources, including satellite, climate, and soil parameters, were combined to predict cotton production over five years from 2017

Table 4.1 : Summary of the data provided with its descriptions, spatial, and temporal resolutions.

Data Source	Features	Spatial Res.	Temporal Res.
MODIS	EVI	463 m	Daily
	LAI & FPAR	500 m	4 days
	LST_D & LST_N	1 km	Daily
SMAP	SSM	10 km	3 days
Daymet V4	P, T_{max} & T_{min} SR, DLD	1 km	Daily
SoilGrids	Sand, Silt, Clay BD, CEC, N, pH	250 m	N/A
USDA	Yield Record	30 m	Annually

to 2021. The satellite-based-generated biophysical parameters directly affecting yield were used to explain the physical and chemical changes in crops during their growth cycle. For this purpose, high spatial and temporal resolution, ready-to-use MODIS products, Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), and Fraction of Photosynthetically Active Radiation (FPAR) were selected. Along with the biophysical parameters, the land surface temperature indicates crop stress and agricultural drought. Hence, day- and night-acquired land surface temperatures (LST_D & LST_N) are used for understanding differences in land surface attributes on the field [136]. Another significant factor affecting crop production is surface soil moisture (SSM), provided worldwide by SMAP satellite in three-day intervals [137]. Due to the high spatial and temporal resolution over CONUS, the climate data were obtained from Daymet V4 [138]. The climate data consists of five variables: precipitation (P), maximum and minimum temperature (T_{max} & T_{min}), solar radiation (SR), and daylight duration (DLD). In addition to these features, another crucial component of agricultural practices is soil property [122]. Soil texture, water holding capacity, and nutrient ingredients of the soil affect crop growth stages and obtained yield. Seven soil properties, which are the sand, silt, clay content of the soil, Bulk density (BD), cation exchange capacity (CEC), total nitrogen (N), and pH, were gathered from [139]. All these data were collected from the Google Earth Engine platform.

4.2.2 Explainable boosting machines

The EBM algorithm as a glass-box model was first introduced by [131] to be as accurate as other boosting-based machine learning algorithms while providing interpretable results to demonstrate a clear understanding of the model for the decision-making process. The EBM is built on the Generalized Additive Models (GAMs) framework, which allows for flexible modeling of the relationship between the dependent variable and multiple independent variables:

$$g(E[y]) = \beta_0 + \sum_{i=1}^{n} f_i(x_i), \tag{4.1}$$

where β_0 is the intercept, f_i is each feature function, n feature dimension, x_i represents the features, and g is the link function that can be used for both regression or classification tasks. This model's major disadvantage is being restricted to univariate terms, which means interactions between features are not considered in the model representation. To overcome this limitation of conventional GAMs, [140] modified the equation by adding pairwise interaction of input features and increased the accuracy of the model while still preserving the intelligibility. The resulting model is called Generalized Additive Models plus Interactions (GA^{2}Ms):

$$g(E[y]) = \beta_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{i,j}(x_i, x_j), \tag{4.2}$$

where $f_{i,j}$ indicates the pairwise interactions. The EBM model is a modified form of GA^{2}Ms that aims to reduce the computational complexity introduced by including pairwise interaction terms. While training the EBM model may take a bit longer compared to similar methods, the prediction stage of the model is faster. For more information on the algorithm and its implementation, please refer to [131] and [140].

4.3 Experimental Study

4.3.1 Data preprocess

This study considered the top cotton-producing states in the United States. Most cotton fields in the CONUS are planted at the end of April and harvested at the

beginning of November; hence, this period was used to limit the temporal data. The crop map data were downsampled to match the spatial resolution of each data source, and cotton masks were used to filter out non-cotton pixels for each year independently. The dynamic features, including MODIS-based products and SMAP SSM, were interpolated to fill the temporal gaps and obtain the exact temporal resolution as the climate data. Afterward, each dynamic feature was first smoothed with a moving average filter to reduce noise and then temporally aggregated to 28 days intervals from April 21^{st} to November 6^{th} with a total of 8 periods, which means 88 dynamic feature dimensions. It must be noted that while the temporal mean for each dynamic feature was calculated, the cumulative sum for precipitation was applied for temporal aggregation. The static features were generated by averaging each pixel's soil properties in different depths. Lastly, the dynamic and static features were concatenated to obtain the final dataset containing 95 dimensions in feature space.

4.3.2 Experimental setup

The performance of the EBM was compared with three standard glass-box methods: ridge linear regression (RLR), least absolute shrinkage and selection operator (LASSO), and decision tree (DT), and two representative boosting-based black-box methods: XGBoost and LightGBM. The entire dataset comprised 5-year-long yield records from 214 counties. This indicates that there are 1070 observations total in the dataset for regression analysis. The dataset was divided randomly into training and testing sets, with 80% and 20% of the data, respectively. To fine-tune the hyperparameters of each method, 5-fold cross-validation was applied to the validation dataset. A Bayesian optimization-based framework, Optuna, was used for tuning the hyperparameters of each method [141]. Optuna tuned all hyperparameters for each method with 500 runs to ensure a fair comparison. The hyperparameters of maximum depth, features, leaf nodes, and minimum samples leaf were evaluated for tree-based methods, including DT, EBM, LightGBM, and XGBoost. In contrast, the regularization term, α value, was optimized for traditional regression methods like RLR and LASSO. Four accuracy metrics, namely mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient

Table 4.2 : Accuracy metrics of the six methods on the test dataset with tuned hyperparameters.

Methods	Models	MAE	RMSE	MAPE	R^2
Glass-box	RLR	110.56	143.29	0.14	0.60
	LASSO	96.04	130.76	0.12	0.67
	DT	94.94	130.00	0.12	0.67
	EBM	87.41	118.57	0.10	0.73
Black-box	XGB	89.01	117.73	0.11	0.73
	LightGBM	86.98	114.27	0.11	0.74

of determination (R^2), were utilized to evaluate the performance of the regression models.

4.4 Results & Discussions

The first part of this study compared the prediction performance of the EBM method to traditional regression models (RLR and LASSO) and tree-based models (DT, XGBoost, and LightGBM) in terms of accuracy for predicting cotton yield. Table 4.2 presents performances of all the methods on the test dataset in terms of several metrics, including MAE, RMSE, MAPE, and R^2, for the end-of-season prediction of cotton yield, with hyperparameters optimized by the cross-validation method.

Based on the results presented in Table 4.2, among the glass-box models, the EBM achieves the lowest MAE and RMSE values (87.41 and 118.57, respectively), indicating better yield predictions against the recorded yield data than RLR, LASSO, and DT. The EBM model also shows the highest R^2 value of 0.73, indicating that it can explain significant variation in the data. Among the black-box models, the LightGBM model appears to have better predictive performance with the lowest MAE and RMSE values of 86.98 and 114.27, respectively. Furthermore, the LightGBM model has the highest R^2 value of 0.74. It is worth noting that although the performance of the LightGBM model is slightly better than the EBM model, the difference is not significant. The findings generally indicate that both EBM and LightGBM models appropriately predict the yield. Nevertheless, it is worth mentioning that the EBM is a glass-box model that does not necessitate any post-hoc technique to clarify its decisions examining yield prediction. On the other hand, the scatter plots, presented in Figure

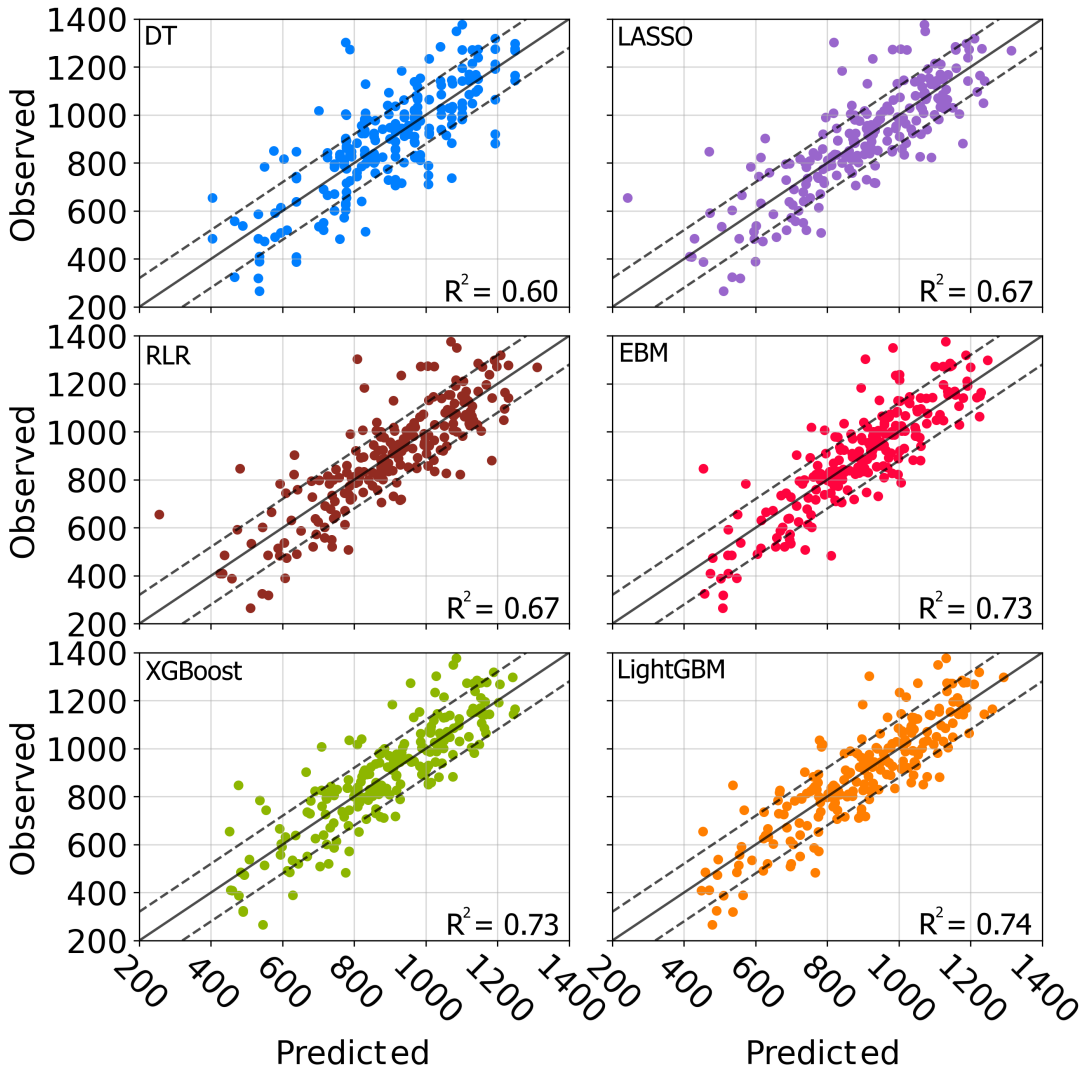


Figure 4.2 : Scatter plot of the predicted versus observed yield data.

4.2, confirm the superiority of the boosted-tree-based models for which the number of predictions that exceed the 10% tolerance limit (dashed lines) are noticeably smaller compared to other methods.

Once the functional relationship between input features and the yield data is established, each feature's particular importance and interactions can be interpreted using the coefficients of the EBM model. The contribution of dynamic features, static features, and feature interactions to the overall feature importance was found as 78%, 6%, and 16%, respectively (see Figure 4.3 & 4.4). The P was the most effective parameter among the others, with an explained importance of 12%, particularly in June, July, and August, followed by EVI with an importance of 9% within the same months. Besides precipitation and EVI, the LAI and the FPAR have the same

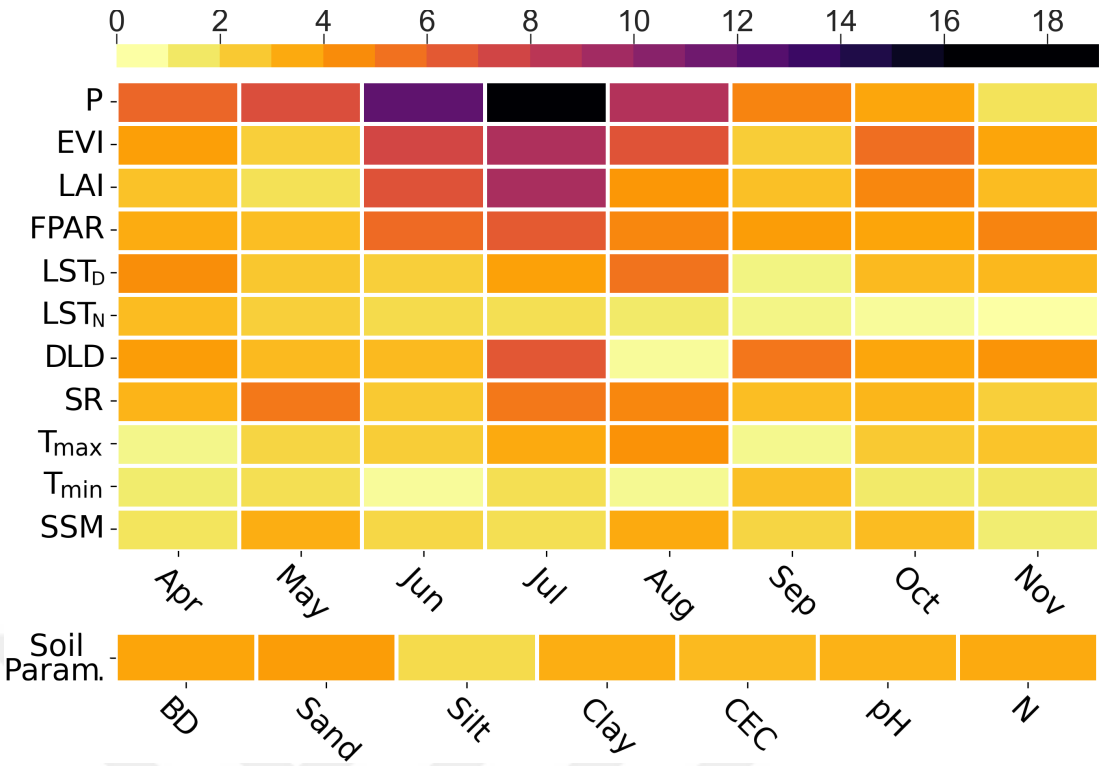


Figure 4.3 : Global importance of dynamic features for each period (above) and static features (below).

importance at around 8%, with the most important periods from June to August as precipitation and EVI. These findings indicate that the greenness of the cotton and its growth represented via biophysical parameters with precipitation rate significantly affect the harvested cotton product during the boll-setting stage, which is in line with [118]. The DLD has shown its significance in July and August, with 7% in overall total importance of features. While the LST_D is effective between April and September, LST_N loses its importance towards harvest. Like LST_D, SR mainly contributes to the model until the beginning of September, from the seeding to the boll-setting stage. T_{max} and T_{min} were found to be less effective dynamic features; however, the sum of the importance of T_{max} and T_{min} corresponds 10% of the overall importance. Along the growing cotton cycle, the SSM was shown unexpectedly low importance, around 6% in total. Its contribution may be suppressed due to the dominance of precipitation.

The distinguishing characteristic of EBM, among other methods, is that it allows the quantification of interactions between input features and the importance of these interactions in understanding the relationship between the predictors and the cotton yield. Sand, silt, soil clay contents, and bulk density directly affect

agricultural productivity because of their relationship with water-holding capacity and air circulation within the soil. For instance, soils with relatively higher silt and clay contents have the more water-holding ability and are more suitable for agriculture. As shown in Figure 4.4, the importance of the bulk&clay, which are inversely proportional, interaction alone accounts for 2% of the global total feature importance. In contrast, individual feature importance of static features accounts for 6% of total feature importance. The significance of the interaction of soil texture with P and EVI in July was found in significant interaction terms, which proves the relation of soil texture with agricultural productivity and climatic factors. Another vital interaction between LST_D in the boll-setting stage and LAI right before the harvesting stage was found relevant. It can be concluded that LST_D in mid-season impacts crop growth, affecting the LAI near the harvesting period.

To evaluate the within-season prediction performance of the EBM, the cotton yield prediction was carried out from the first month to the last, adding each month to the model cumulatively. Figure 4.5 shows the accuracy of each model by RMSE and R^2 metrics. The accuracy achieved by using features from April to August was 0.7, corresponding to $\sim$95% of the accuracy depicted by the end-of-season model. This result aligns with Figure 4.3, where most essential dynamic features lie for the end-of-season prediction model between April and August. It can be concluded from the corporation of Figure 4.5 and Figure 4.3 that the EBM enables an accurate and reliable regression model for within-season prediction of cotton yield.

4.5 Conclusion

In this research, cotton yield prediction has been investigated to obtain an accurate and explainable model with the integration of satellite remote sensing images, climate data, and soil parameters by applying the state-of-art EBM method. The study's findings proved that EBM as a glass-box method showed more accurate results than other glass-box methods and had comparable results with widely used black-box methods, XGBoost and LightGBM. The interactions between the features, their importance, and their interpretation were determined with EBM during the growth cycle of the cotton without applying any further post-hoc explanation methods. The explainability

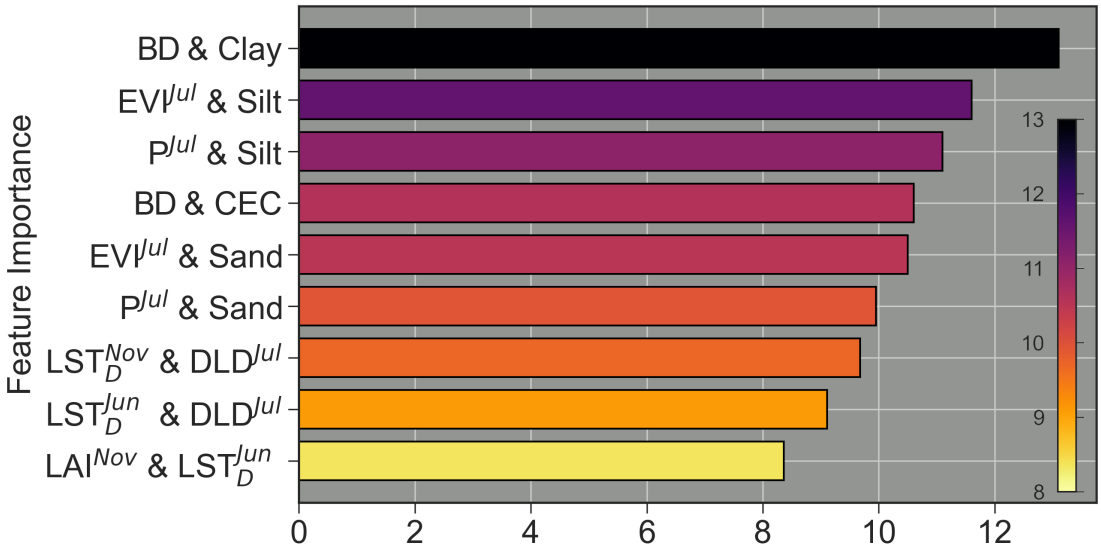


Figure 4.4 : Global importance of interactions between features where the sum of the interaction importance corresponds to 16% of the global importance.

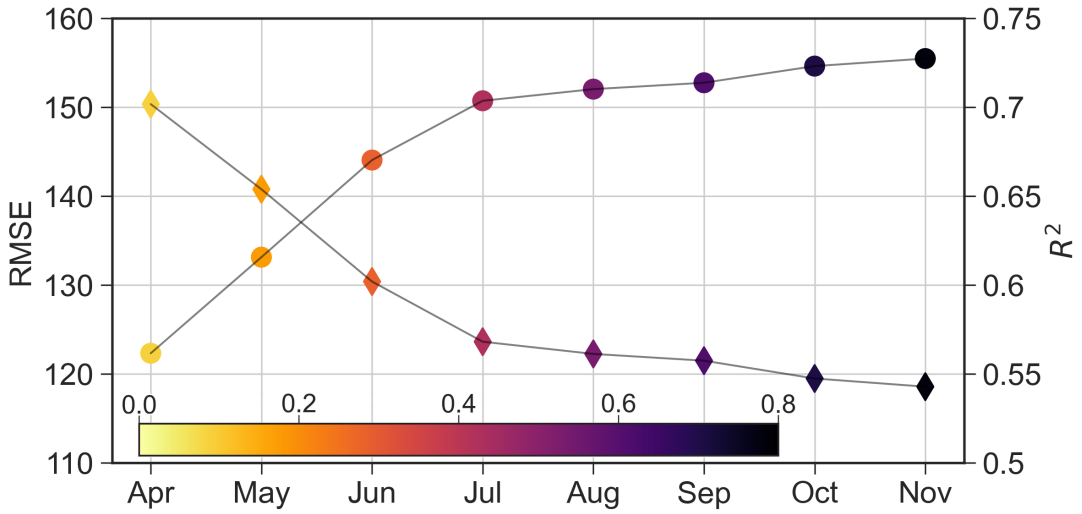


Figure 4.5 : Within season cotton yield prediction accuracy by months. The color bar indicates the cumulative feature importance during the growth cycle of cotton.

of the EBM revealed that P, EVI, LAI, and FPAR until mid-season are the driving factors to predict the cotton yield. The LST_D, DLD, SR, and SSM appear less important than the P and biophysical parameters; however, they contributed to the model to improve the prediction accuracy in certain months along the growth cycle. The importance of T_{min} and LST_N does not fluctuate in the temporal domain and shows relatively more minor contributions; nevertheless, their total feature importance should not be left aside. Besides the dynamic features, static features and feature interactions significantly affect the model prediction performance. Furthermore, the monthly-based implementation of EBM has shown that cotton yield estimation can

be achieved four months in advance with high accuracy. In future work, the study will be extended to yield prediction of different agricultural products by taking into account the transferring capability of the EBM model in other geographic locations where climatic factors and soil properties vary.

5. CONCLUSIONS

The availability of agricultural fields is diminishing as the global population continues to increase. Food security is further complicated by the threat of climate change and soil degradation. However, the utilization of remote sensing satellite imagery presents a valuable solution for monitoring agricultural fields and gathering essential data on crop health, growth stages, soil moisture, and potential yield. This data, in turn, can be leveraged to enhance agricultural practices and production. Satellite data analysis can uncover significant patterns that aid decision-makers in making informed choices regarding agricultural management by employing AI methods. These methods can also illuminate the complex relationships between various factors and crop yield. With the help of this information, farmers can improve their yields while minimizing their environmental impact. Through the integration of satellite data and AI, we can effectively contribute to ensuring food security for the future.

The first chapter of this thesis focuses on two crucial biophysical parameters: LAI and NDVI, both estimated from full polarimetric Synthetic Aperture Radar (SAR) images. The study utilizes data from the AgriSAR2009 campaign funded by the ESA, specifically focusing on the Indian Head region. The data set covers eight fields with four crop types: field pea, barley, canola, and oat. RADARSAT-2 fully polarimetric images were acquired during a single growth period from June to July, along with in-situ measurements of LAI and NDVI collected at the corners of each field during the corresponding growing cycle. The IDW interpolation method was employed for the NDVI and LAI in-situ measurements to determine the output value for each pixel. A total of 23 polarimetric features were derived from fully polarimetric RADARSAT-2 images to investigate the relationship with biophysical parameters. Since GSA with Sobol's theorem needed uncorrelated input feature space, NCA was implemented in the data preprocessing step to create a sub-feature space. The uncorrelated polarimetric features were selected for each crop and biophysical parameter. The PCE

method is a regression technique that establishes the functional relationship between inputs and outputs using orthogonal polynomials, considering the interaction between higher-order polynomials of input parameters, enhancing the regression performance. A further advantage of the PCE is that the polynomial coefficient can be used to carry out the GSA in accordance with Sobol's theorem. The main findings of the first research could be summarized as follows:

- Some of the derived polarimetric features are highly correlated with each other. First, principal component analysis was applied in [27] to create an uncorrelated and independent input space. However, transforming the input space into another space prevents the investigation of the effective polarimetric features for the estimation of LAI and NDVI. Therefore, NCA was implemented to eliminate correlated polarimetric features and to create a sub-feature space while preserving the original feature space.

- An infinite series can be constructed with orthogonal polynomials and full interactions. The degree of polynomials and interactions should be truncated to effectively solve the coefficients of polynomials. As stated in [36], more than triple interactions and high-order polynomial interactions have no significant effect on the regression model while drastically increasing the coefficient of polynomials. For this reason, the constructed PCE was limited to maximum triple interaction and up to 15^{th} degree polynomial. The PCE results showed that beyond the 10^{th} degree of polynomial expansion does not contribute significantly to the estimation and creates over-fitting of the regression model while increasing the computational effort.

- According to the regression results of each crop type model, LAI estimation accuracy varies between 0.90 and 0.99 in terms of R^2, and the corresponding $RMSE$ values fall between $0.25 - 0.45$, and NDVI estimation accuracy varies between 0.90 and 0.99 in terms of R^2, and the corresponding $RMSE$ values fall between $0.03 - 0.08$. The regression accuracy of Case-I and Case-II demonstrated that it is insignificant evaluating each field individually. Accuracy assessments show that the

NDVI parameter can be slightly better estimated compared to LAI, especially for field pea and canola, followed by oat and barley.

- Based on the overall GSA results, it can be concluded that up to 2^{nd} order interactions between polarimetric features dominate the regression model of the NDVI parameter, while the 3^{rd} order interactions were found to be still effective for LAI estimation, which affected the regression model more than %5 for each crop type.

- The GSA of field pea showed that the polarimetric features, $F4 : |VV|^2$ & $F11 : \rho_{HH,VV}$, are the driven features for the estimation of LAI, and the model performance reached 0.82 and 0.48 in terms of R^2 and $RMSE$, respectively. For the estimation of NDVI, using the polarimetric feature $F3 : |HV|^2$ alone, the $R^2 : 0.93$ can be achieved with the PCE regression model. With the addition of the $F4$ polarimetric feature and the involvement of the interaction between $F3 : |HV|^2$ & $F4 : |VV|^2$, the model accuracy reached 0.98 and 0.035 in terms of R^2 and RMSE, respectively.

- The regression model of the barley for LAI estimation is mostly driven by $F11 : \rho_{HH,VV}$ with an accuracy of $R^2 = 0.85$ and $RMSE = 0.47$. The model accuracy can be improved up to $R^2 = 0.89$ and $RMSE = 0.68$ by adding $F2 : |HH|^2$ and $F6 : A$ polarimetric features. By using $F2 : |HH|^2$ together with $F4 : |VV|^2$, which is quite effective when used individually, the NDVI parameter for barley can be estimated with a model accuracy of $R^2 = 0.90$ and $RMSE = 0.082$.

- The most effective polarimetric feature for canola was found to be $F3 : |HV|^2$ for both LAI and NDVI estimation. The model accuracy was higher than $R^2 = 0.85$, and by adding the $F2 : |HH|^2$ and its interaction with $F3 : |HV|^2$, the model accuracy exceeded $R^2 = 0.90$ for LAI estimation. However, using $F3 : |HV|^2$ alone, the estimation accuracy of NDVI was reached to $R^2 = 0.94$ and $RMSE = 0.027$.

- Although $F3 : |HV|^2$ was found to be the most effective polarimetric feature in estimating the LAI and NDVI of oat, three more polarimetric features, which are $F2 : |HH|^2$, $F14 : \phi_{HH,VV}$, and $F20 : \rho_{HH+VV,HH-VV}$ should be used in order to

reach $R^2 : 0.90$ and $RMSE : 0.3$ of the model accuracy for LAI. In the case of NDVI estimation, by adding $F2 : |HH|^2$ and $F20 : \rho_{HH+VV,HH-VV}$ polarimetric features to the model, the accuracy of $R^2 = 0.90$ and $RMSE = 0.08$ can be achieved.

- The findings of this study demonstrate that the most relevant polarimetric features for NDVI estimation are the backscatter coefficients. In addition to the backscatter coefficients, the $F11 : \rho_{HH,VV}$ polarimetric feature was discovered to be effective for the estimation of the LAI biophysical parameter of field pea and barley.

In the second chapter of this thesis, the focus is on estimating soil moisture, a critical biophysical parameter that profoundly impacts crop growth and health. Monitoring soil moisture can be achieved using SAR satellite missions, which are sensitive to soil water content due to their effect on the dielectric constant. Two widely used satellite missions for monitoring soil moisture are SMAP, providing quasi-global coverage with high temporal and low spatial resolutions, and Sentinel-1 missions with high spatial and lower temporal resolutions. However, neither the high spatial resolution of Sentinel-1 nor the high temporal resolution of SMAP can offer daily soil moisture data suitable for field-scale agricultural activities, especially irrigation-related. Therefore, accurate SM prediction addresses this limitation and supports effective agricultural practices. This study employs time series analysis by utilizing the LSTM method to achieve accurate and daily field-scale soil moisture predictions. The LSTM approach leverages the sequential data provided by the high temporal resolution SMAP SM product, high spatial resolution Sentinel-1 backscatter coefficients, climate and topographic data, and soil texture relevant to agricultural activities. The LSTM model was trained using - microwave radar datasets from SMAP and Sentinel-1, as well as soil texture, climate, and topographical data, which serve as predictors for soil moisture and ground station data from the ISMN located worldwide. By combining microwave remote sensing satellite observations and auxiliary data, the created model successfully achieves short-term soil moisture forecasts at the field scale. The key findings from the second phase of the research can be summarized as follows:

- The grid search algorithm was applied by using various hidden layers and unit sizes, learning rates, loss functions, and optimization functions for hyperparameter

optimization. The performances of the trained models with setups having four different window sizes varying from 1 day to 30 days were examined. The best accuracy metrics were achieved for the 5-days window size with $R^2 = 0.87$ and $MAE = 0.033$.

- The model performance was tested with the different data sources adding one-by-one. The evaluation of the models started with climatic factors as the main driver of the SM, and soil texture, topography, SMAP, and S1 data were added one by one, respectively. The model trained solely with climatic factors achieved $R^2 = 0.34$ and $MAE = 0.085$. However, the accuracy assessment criteria significantly improved as each data source was incorporated. With the inclusion of static parameters, the model accuracy increased to $R^2 = 0.84$ and $MAE = 0.037$. Subsequently, by integrating SMAP SM data, the accuracy further improved to $R^2 = 0.86$ and $MAE = 0.034$. Lastly, the addition of S1 data resulted in a slight additional increase in model accuracy.

- The accuracy of the optimized LSTM model was found to be effective for SM prediction with the R^2 of 0.87, *RMSE* of 0.046, *ubRMSE* of 0.045, and *MAE* of 0.033. The model accurately predicts soil moisture values for the following day, offering high spatial resolution across regions with various geophysical properties and climate classifications.

- The physical characteristics of the land cover influence the prediction accuracy of the developed LSTM model. The highest model accuracy was achieved for the shrubland LCC with an MAE of 0.02 and followed by cropland, grassland, and tree classes with an MAE of 0.03. The complexity of the surface reduces the performance of the model, which is the case for mosaic LCC which the model has the capability to estimate the SM with MAE around 0.05.

- Another finding showed that the NDVI variation is one of the reasons for the deterioration of the SM prediction. The vegetation coverage, especially in cropland, and the backscatter signals of SAR data are strongly affected by biomass due to the interaction between signals, crops, and soil. The model performance was observed

around an $MAE = 0.1$ for low-vegetated areas, while the presence of vegetation decreased the accuracy of the model, varying from 0.2 to 0.4 in terms of MAE.

- The water holding capacity of the soil is directly related to particle size of soil texture. The soil that has more clay content tends to preserve SM, while the higher sand content causes air accumulation within the soil that prevents the water from holding into the soil. The study revealed that 61% of the stations had sandy soil texture, with the model achieving a performance of approximately $MAE = 0.03$. In contrast, among the remaining stations, 38% had silty soil, and the model's performance experienced a slight decrease to around $MAE = 0.04$. Another reason for decreasing model performance in silty soil is that these types of soils are more suitable for agricultural practices, which also affects the model because of vegetation coverage.

- The model performance was evaluated according to the climate classes of the stations used in the study. The climatic factors are the main drivers for the SM prediction, especially precipitation is a significant factor that negatively impacts the SM prediction due to the change in the interaction between SAR signals and land surface. The temperate (C) climate class receives higher precipitation than the dry (B) class. Moreover, when considering subclasses, arid climates (Bw) experience more precipitation than semi-arid climates (Bs), and no dry season (Cf) receives more precipitation than dry summer (Cs). The model performance can be arranged in ascending order of MAE as follows: Bw, Bs, Cs, and Cf, respectively.

The third chapter of this thesis focuses on yield estimation for cotton to understand the spatial and temporal variations of cotton yield and their relationship with changes in climatic and soil conditions is challenging yet essential for making efficient agricultural decisions. Developing a reliable yield estimation model that aids farmers in agricultural planning requires a clear interpretation of the functional relationship between biophysical parameters, climate data, soil properties, and cotton yield. The CONUS was selected as the study area due to the comprehensive investigation of cotton production by the USDA, which annually releases a high-resolution crop classification map with yield records. The integration of MODIS products EVI, LAI,

and FPAR, as well as LST_D and LST_N, surface SM data from SMAP, climate data from Daymet V4 such as precipitation, air temperature, solar radiation, and daylight duration) and soil properties from SoilGrid, such as texture, bulk density, and nutrient content, were considered as the predictors for cotton. The aim of incorporating these diverse features was to establish a comprehensive understanding of the functional relationship between yield records and their associated factors. The EBM, which considers pairwise interactions between input vectors, was implemented to provide interpretable and explainable results without the need for post-hoc methods. The analysis revealed the significance of specific biophysical and climatic parameters during different phenological stages and identified soil properties that influenced cotton yield within the model. The main discoveries from the third phase of the research can be outlined as follows:

- The dynamic features, which include MODIS-based products and SMAP SSM, were interpolated to fill the temporal gaps and match the temporal resolution with the climate data. Subsequently, a two-step process was applied to each dynamic feature. Firstly, it was smoothed using a moving average filter to reduce noise. Then, the features were temporally aggregated into 28-day intervals from April to November, comprising a total of 8 periods, resulting in 88 dynamic feature dimensions. Finally, the dynamic and static features were combined, resulting in a dataset comprising 95 dimensions in the feature space. The entire dataset comprised 5-year-long yield records from 214 counties. This indicates that there are 1070 observations total in the dataset for regression analysis.

- The dataset was divided randomly 80%-20% as a training and testing set. For hyperparameter tuning, 5-fold cross-validation was carried out with a Bayesian optimization framework, namely Optuna, for each method. According to accuracy analysis, the EBM as a glass-box method showed comparable accuracy against the black-box methods with the *MAE* of 87.41 lb/acre, *RMSE* of 118.57 lb/acre, *MAPE* of %10, and R^2 of 0.73.

- The individual effect of dynamic features was examined, and it was found that the importance of dynamic features corresponds to 78% of the total importance score

of all features and their interactions, followed by static features and interactions as 6% and 12%, respectively.

- Since cotton is mainly produced in rain-fed agricultural fields, precipitation was found to be the most effective feature for yield prediction among the others with 12% importance proportional to total feature importance and suppressed the importance of SM. The other important drivers for cotton yield were found as EVI, LAI, and FPAR with the importance of 9%, 8%, and 8%, respectively. These findings proved that with the aid of precipitation, cotton greenness along with the physical improvement during the growth cycle is an important indicator for obtained yield.

- Even though temperature related parameters such as T_{max}, T_{min}, LST_D and LST_N seem to be less important compared to precipitation and biophysical parameters, the total importance of these parameters corresponds to 23% of total feature importance. This finding reveals the effect of air and surface temperature on the development of cotton during the phenological cycle.

- Although the importance of individual importance of static features was found 6%, the interactions between static features and with EVI and P, especially in July, highlight the importance of static features.

- By determining the importance of temporal features on a monthly basis, it is possible to analyze which parameters are effective during the phenological cycle of cotton and to determine the amount of yield to be obtained. The majority of the important features are gathered in the first four months of the growth cycle. The feature importance and monthly basis regression analysis proved that the obtained yield can be predicted in mid-season with high accuracy.

The use of explainable artificial intelligence (XAI) models in agricultural research has ensured valuable insights into the phenological cycle and provided reliable accuracy in estimating target parameters such as biophysical parameters and yield. Incorporating XAI models has achieved a milestone in agricultural studies, as they offer transparent interpretations of effective predictors. Nevertheless, when dealing

with satellite remote sensing data and derived parameters, one must consider the inherent uncertainty stemming from measurement error and heterogeneity in the target domain. For agricultural phenomena, this uncertainty in high-dimensional input parameters naturally affects the accuracy of the estimated target parameter. Future research will focus deeper on understanding and quantifying the uncertainties present within XAI models, aiming to enhance the accuracy and reliability of estimations for the target parameter, paving the way for even more robust agricultural practices.

REFERENCES

[1] **United Nations Department of Economic and Social Affairs, Population Division** (2022). *World population prospects 2022: Summary of results*, UN DESA/POP/2022/TR/NO. 3.

[2] **World Bank** (2023). *Agricultural land (% of land area)*, retrieved from: https://data.worldbank.org/indicator/AG.LND.AGRI.ZS.

[3] **Mandal, D.**, **Bhattacharya, A. and Rao, Y.S.** (2021). Biophysical parameter retrieval using full- and dual-pol SAR data, Springer Singapore, Singapore, pp.107–153.

[4] **Zhang, C.**, **Marzougui, A. and Sankaran, S.** (2020). High-resolution satellite imagery applications in crop phenotyping: An overview, *Computers and Electronics in Agriculture*, *175*, 105584.

[5] **Steele-Dunne, S.C.**, **McNairn, H.**, **Monsivais-Huertero, A.**, **Judge, J.**, **Liu, P.W. and Papathanassiou, K.** (2017). Radar remote sensing of agricultural canopies: a review, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(5), 2249–2273.

[6] **Nasirzadehdizaji, R.**, **Cakir, Z.**, **Balik Sanli, F.**, **Abdikan, S.**, **Pepe, A. and Calò, F.** (2021). Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring, *Computers and Electronics in Agriculture*, *185*, 106118.

[7] **Sekertekin, A.**, **Marangoz, A.M. and Abdikan, S.** (2020). ALOS-2 and Sentinel-1 SAR data sensitivity analysis to surface soil moisture over bare and vegetated agricultural fields, *Computers and Electronics in Agriculture*, *171*, 105303.

[8] **Yuzugullu, O.**, **Erten, E. and Hajnsek, I.** (2018). Assessment of paddy rice height: sequential inversion of coherent and incoherent models, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(9), 3001–3013.

[9] **Wu, S.**, **Yang, P.**, **Ren, J.**, **Chen, Z.**, **Liu, C. and Li, H.** (2020). Winter wheat LAI inversion considering morphological characteristics at different growth stages coupled with microwave scattering model and canopy simulation model, *Remote Sensing of Environment*, *240*, 111681.

[10] **Inoue, Y.**, **Sakaiya, E. and Wang, C.** (2014). Capability of C-band backscattering coefficients from high-resolution satellite SAR sensors to assess biophysical variables in paddy rice, *Remote Sensing of Environment*, *140*, 257–266.

[11] **Macelloni, G.**, **Paloscia, S.**, **Pampaloni, P.**, **Marliani, F. and Gai, M.** (2001). The relationship between the backscattering coefficient and the biomass of narrow and broad leaf crops, *IEEE Transactions on Geoscience and Remote Sensing*, *39*(4), 873–884.

[12] **Schlund, M. and Erasmi, S.** (2020). Sentinel-1 time series data for monitoring the phenology of winter wheat, *Remote Sensing of Environment*, *246*, 111814.

[13] **Silva-Perez, C.**, **Marino, A.**, **Lopez-Sanchez, J.M. and Cameron, I.** (2021). Multitemporal polarimetric SAR change detection for crop monitoring and crop type classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 12361–12374.

[14] **Mascolo, L.**, **Lopez-Sanchez, J.M.**, **Vicente-Guijalba, F.**, **Nunziata, F.**, **Migliaccio, M. and Mazzarella, G.** (2016). A complete procedure for crop phenology estimation with PolSAR data based on the complex wishart classifier, *IEEE Transactions on Geoscience and Remote Sensing*, *54*(11), 6505–6515.

[15] **Hariharan, S.**, **Mandal, D.**, **Tirodkar, S.**, **Kumar, V.**, **Bhattacharya, A. and Lopez-Sanchez, J.M.** (2018). A novel phenology based feature subset selection technique using random forest for multitemporal PolSAR crop classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(11), 4244–4258.

[16] **Wang, H.**, **Magagi, R.**, **Goïta, K.**, **Trudel, M.**, **McNairn, H. and Powers, J.** (2019). Crop phenology retrieval via polarimetric SAR decomposition and random forest algorithm, *Remote Sensing of Environment*, *231*, 111234.

[17] **Vicente-Guijalba, F.**, **Martinez-Marin, T. and Lopez-Sanchez, J.M.** (2015). Dynamical approach for real-time monitoring of agricultural crops, *IEEE Transactions on Geoscience and Remote Sensing*, *53*(6), 3278–3293.

[18] **Canisius, F.**, **Shang, J.**, **Liu, J.**, **Huang, X.**, **Ma, B.**, **Jiao, X.**, **Geng, X.**, **Kovacs, J.M. and Walters, D.** (2018). Tracking crop phenological development using multi-temporal polarimetric RADARSAT-2 data, *Remote Sensing of Environment*, *210*, 508–518.

[19] **Mandal, D.**, **Kumar, V.**, **Ratha, D.**, **Dey, S.**, **Bhattacharya, A.**, **Lopez-Sanchez, J.M.**, **McNairn, H. and Rao, Y.S.** (2020). Dual polarimetric radar vegetation index for crop growth monitoring using Sentinel-1 SAR data, *Remote Sensing of Environment*, *247*, 111954.

[20] **Lopez-Sanchez, J.M.**, **Cloude, S.R. and Ballester-Berman, J.D.** (2012). Rice phenology monitoring by means of SAR polarimetry at X-Band, *IEEE Transactions on Geoscience and Remote Sensing*, *50*(7), 2695–2709.

[21] **Lopez-Sanchez, J.M.**, **Vicente-Guijalba, F.**, **Ballester-Berman, J.D. and Cloude, S.R.** (2014). Polarimetric response of rice fields at C-Band: analysis and phenology retrieval, *IEEE Transactions on Geoscience and Remote Sensing*, *52*(5), 2977–2993.

[22] **Erten, E.**, **Taşkin, G. and Lopez-Sanchez, J.M.** (2019). Selection of PolSAR observables for crop biophysical variable estimation with global sensitivity analysis, *IEEE Geoscience and Remote Sensing Letters*, *16*(5), 766–770.

[23] **Hariri-Ardebili, M.A. and Sudret, B.** (2020). Polynomial chaos expansion for uncertainty quantification of dam engineering problems, *Engineering Structures*, *203*, 109631.

[24] **Skarbeli, A.V. and Álvarez-Velarde, F.** (2020). Sparse polynomial chaos expansion for advanced nuclear fuel cycle sensitivity analysis, *Annals of Nuclear Energy*, *142*, 107430.

[25] **Yuzugullu, O.**, **Erten, E. and Hajnsek, I.** (2017). A multi-year study on rice morphological parameter estimation with X-Band PolSAR data, *Applied Sciences*, *7*(6).

[26] **Yuzugullu, O.**, **Marelli, S.**, **Erten, E.**, **Sudret, B. and Hajnsek, I.** (2015). Global sensitivity analysis of a morphology based electromagnetic scattering model, *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp.2743–2746.

[27] **Celik, M.F. and Erten, E.** (2021). Principal component analysis based polynomial chaos expansion regression of leaf area index from PolSAR imagery, *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp.6096–6099.

[28] **Torre, E.**, **Marelli, S.**, **Embrechts, P. and Sudret, B.** (2019). Data-driven polynomial chaos expansion for machine learning regression, *Journal of Computational Physics*, *388*, 601–623.

[29] **Oladyshkin, S. and Nowak, W.** (2012). Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, *Reliability Engineering System Safety*, *106*, 179–190.

[30] **Sudret, B.** (2008). Global sensitivity analysis using polynomial chaos expansions, *Reliability engineering & system safety*, *93*(7), 964–979.

[31] **Crestaux, T.**, **Le Maıtre, O. and Martinez, J.M.** (2009). Polynomial chaos expansion for sensitivity analysis, *Reliability Engineering & System Safety*, *94*(7), 1161–1172.

[32] **Rosenblatt, M.** (1952). Remarks on a multivariate transformation, *The annals of mathematical statistics*, *23*(3), 470–472.

[33] **Nataf, A.** (1962). Détermination des distributions de probabilités dont les marges sont données, *C. R. Acad. Sci., Paris*, *255*, 42–43.

[34] **Mara, T.A. and Becker, W.E.** (2021). Polynomial chaos expansion for sensitivity analysis of model output with dependent inputs, *Reliability Engineering & System Safety*, 107795.

[35] **Caves, R.**, **Davidson, G.**, **Padda, J. and Ma, A.** (2011). AgriSAR2009 final report, Vol. 1. Executive summary, data acquisition, data simulation, *ESA, Paris, France, Tech. Rep*.

[36] **Blatman, G. and Sudret, B.** (2011). Adaptive sparse polynomial chaos expansion based on least angle regression, *Journal of Computational Physics*, *230*(6), 2345–2367.

[37] **Efron, B.**, **Hastie, T.**, **Johnstone, I. and Tibshirani, R.** (2004). Least angle regression, *The Annals of statistics*, *32*(2), 407–499.

[38] **Marelli, S.**, **Lüthen, N. and Sudret, B.** (2021). UQLab user manual – Polynomial chaos expansions, **Technical Report**, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, report n UQLab-V1.4-104.

[39] **Liu, C.**, **Shang, J.**, **Vachon, P.W. and McNairn, H.** (2013). Multiyear crop monitoring using polarimetric RADARSAT-2 data, *IEEE Transactions on Geoscience and Remote Sensing*, *51*(4), 2227–2240.

[40] **Goldberger, J.**, **Hinton, G.E.**, **Roweis, S. and Salakhutdinov, R.R.** (2004). Neighbourhood components analysis, *Advances in neural information processing systems*, *17*.

[41] **Jung, H.C.**, **Kang, D.H.**, **Kim, E.**, **Getirana, A.**, **Yoon, Y.**, **Kumar, S.**, **Peters-lidard, C.D. and Hwang, E.** (2020). Towards a soil moisture drought monitoring system for South Korea, *Journal of Hydrology*, *589*, 125176.

[42] **Berg, A. and Sheffield, J.** (2018). Climate change and drought: the soil moisture perspective, *Current Climate Change Reports*, *4*(2), 180–191.

[43] **Norbiato, D.**, **Borga, M.**, **Degli Esposti, S.**, **Gaume, E. and Anquetin, S.** (2008). Flash flood warning based on rainfall thresholds and soil moisture conditions: an assessment for gauged and ungauged basins, *Journal of Hydrology*, *362*(3-4), 274–290.

[44] **Martínez-Fernández, J.**, **González-Zamora, A.**, **Sánchez, N.**, **Gumuzzio, A. and Herrero-Jiménez, C.** (2016). Satellite soil moisture for agricultural drought monitoring: assessment of the SMOS derived soil water deficit index, *Remote Sensing of Environment*, *177*, 277–286.

[45] **Efremova, N.**, **Seddik, M.E.A. and Erten, E.** (2021). Soil moisture estimation using Sentinel-1/-2 imagery coupled with cycleGAN for time-series gap filing, *IEEE Transactions on Geoscience and Remote Sensing*.

[46] **Lawless, C.**, **Semenov, M.A. and Jamieson, P.D.** (2008). Quantifying the effect of uncertainty in soil moisture characteristics on plant growth using a crop simulation model, *Field Crops Research*, *106*, 138–147.

[47] **Dai, X.**, **Huo, Z. and Wang, H.** (2011). Simulation for response of crop yield to soil moisture and salinity with artificial neural network, *Field Crops Research*, *121*, 441–449.

[48] **Famiglietti, J.**, **Rudnicki, J. and Rodell, M.** (1998). Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, *Journal of Hydrology*, *210*, 259–281.

[49] **Western, A.W.**, **Grayson, R.B.**, **Blöschl, G.**, **Willgoose, G.R. and McMahon, T.A.** (1999). Observed spatial organization of soil moisture and its relation to terrain indices, *Water Resources Research*, *35*, 797–810.

[50] **Moeslund, J.E.**, **Arge, L.**, **Bøcher, P.K.**, **Dalgaard, T.**, **Odgaard, M.V.**, **Nygaard, B. and Svenning, J.C.** (2013). Topographically controlled soil moisture is the primary driver of local vegetation patterns across a lowland region, *Ecosphere*, *4*, art91.

[51] **Gwak, Y. and Kim, S.** (2017). Factors affecting soil moisture spatial variability for a humid forest hillslope, *Hydrological Processes*, *31*, 431–445.

[52] **Vereecken, H.**, **Kamai, T.**, **Harter, T.**, **Kasteel, R.**, **Hopmans, J. and Vanderborght, J.** (2007). Explaining soil moisture variability as a function of mean soil moisture: A stochastic unsaturated flow perspective, *Geophysical Research Letters*, *34*, L22402.

[53] **Rosenbaum, U.**, **Bogena, H.R.**, **Herbst, M.**, **Huisman, J.A.**, **Peterson, T.J.**, **Weuthen, A.**, **Western, A.W. and Vereecken, H.** (2012). Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale, *Water Resources Research*, *48*, 2011WR011518.

[54] **Wilson, D.J.**, **Western, A.W. and Grayson, R.B.** (2004). Identifying and quantifying sources of variability in temporal and spatial soil moisture observations, *Water Resources Research*, *40*.

[55] **Teuling, A.J.**, **Hupet, F.**, **Uijlenhoet, R. and Troch, P.A.** (2007). Climate variability effects on spatial soil moisture dynamics, *Geophysical Research Letters*, *34*, L06406.

[56] **Wang, T.**, **Franz, T.E.**, **Li, R.**, **You, J.**, **Shulski, M.D. and Ray, C.** (2017). Evaluating climate and soil effects on regional soil moisture spatial variability using EOFs, *Water Resources Research*, *53*, 4022–4035.

[57] **Liu, M.**, **Huang, C.**, **Wang, L.**, **Zhang, Y. and Luo, X.** (2020). Short-term soil moisture forecasting via gaussian process regression with sample selection, *Water*, *12*(11), 3085.

[58] **Yu, J.**, **Zhang, X.**, **Xu, L.**, **Dong, J. and Zhangzhong, L.** (2021). A hybrid CNN-GRU model for predicting soil moisture in maize root zone, *Agricultural Water Management*, *245*, 106649.

[59] **Li, Q.**, **Zhu, Y.**, **Shangguan, W.**, **Wang, X.**, **Li, L. and Yu, F.** (2022). An attention-aware LSTM model for soil moisture and soil temperature prediction, *Geoderma*, *409*, 115651.

[60] **O, S. and Orth, R.** (2021). Global soil moisture data derived through machine learning trained with in-situ measurements, *Scientific Data*, *8*(1), 170.

[61] **Souissi, R.**, **Zribi, M.**, **Corbari, C.**, **Mancini, M.**, **Muddu, S.**, **Tomer, S.K.**, **Upadhyaya, D.B. and Al Bitar, A.** (2022). Integrating process-related information into an artificial neural network for root-zone soil moisture prediction, *Hydrology and Earth System Sciences*, *26*(12), 3263–3297.

[62] **Dobson, M. and Ulaby, F.** (1986). Active microwave soil moisture research, *IEEE Transactions on Geoscience and Remote Sensing*, *GE-24*, 23–36.

[63] **Njoku, E.G. and Entekhabi, D.** (1996). Passive microwave remote sensing of soil moisture, *Journal of Hydrology*, *184*, 101–129.

[64] **Entekhabi, D.**, **Njoku, E.G.**, **O'Neill, P.E.**, **Kellogg, K.H.**, **Crow, W.T.**, **Edelstein, W.N. and Zyl, .J.V.** (2010). The soil moisture active passive (SMAP) mission, *Proceedings of the IEEE*, *98*, 704–716.

[65] **Kerr, Y.H.**, **Waldteufel, P.**, **Richaume, P.**, **Wigneron, J.P.**, **Ferrazzoli, P. and Delwart, A.M..S.** (2012). The SMOS soil moisture retrieval algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, *50*, 1384–1403.

[66] **Sadri, S.**, **Pan, M.**, **Wada, Y.**, **Vergopolan, N.**, **Sheffield, J.**, **Famiglietti, J.S.**, **Kerr, Y. and Wood, E.** (2020). A global near-real-time soil moisture index monitor for food security using integrated SMOS and SMAP, *Remote Sensing of Environment*, *246*, 111864.

[67] **Peng, J.**, **Niesel, J. and Loew, A.** (2015). Evaluation of soil moisture downscaling using a simple thermal-based proxy – the REMEDHUS network (Spain) example, *Hydrology and Earth System Sciences*, *19*, 4765–4782.

[68] **Peng, J.**, **Loew, A.**, **Zhang, S.**, **Wang, J. and Niesel, J.** (2016). Spatial Downscaling of Satellite Soil Moisture Data Using a Vegetation Temperature Condition Index, *IEEE Transactions on Geoscience and Remote Sensing*, *54*, 558–566.

[69] **Hornacek, M.**, **Wagner, W.**, **Sabel, D.**, **Truong, H.L.**, **Snoeij, P.**, **Hahmann, T.**, **Diedrich, E. and Doubkova, M.** (2012). Potential for high resolution systematic global surface soil moisture retrieval via change detection using Sentinel-1, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*, 1303–1311.

[70] **Gao, Q.**, **Zribi, M.**, **Escorihuela, M.J. and Baghdadi, N.** (2017). Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at 100 m resolution, *Sensors*, *17*(9), 1966.

[71] **Liu, Z.**, **Li, P. and Yang, J.** (2017). Soil moisture retrieval and spatiotemporal pattern analysis using Sentinel-1 data of Dahra, Senegal, *Remote Sensing*, *9*, 1197.

[72] **Fan, D.**, **Zhao, T.**, **Jiang, X.**, **Xue, H.**, **Moukomla, S.**, **Kuntiyawichai, K. and Shi, J.** (2022). Soil moisture retrieval from Sentinel-1 time-series data over croplands of Northeastern Thailand, *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

[73] **Nguyen, H.H.**, **Cho, S.**, **Jeong, J. and Choi, M.** (2021). A D-vine copula quantile regression approach for soil moisture retrieval from dual polarimetric SAR Sentinel-1 over vegetated terrains, *Remote Sensing of Environment*, *255*, 112283.

[74] **Attarzadeh, R.**, **Amini, J.**, **Notarnicola, C. and Greifeneder, F.** (2018). Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at plot scale, *Remote Sensing*, *10*, 1285.

[75] **Greifeneder, F.**, **Notarnicola, C. and Wagner, W.** (2021). A machine learning-based approach for surface soil moisture estimations with google earth engine, *Remote Sensing*, *13*(11), 2099.

[76] **Xue, Z.**, **Zhang, Y.**, **Zhang, L. and Li, H.** (2022). Ensemble learning embedded with gaussian process regression for soil moisture estimation: a case study of the continental U.S., *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–17.

[77] **Lei, F.**, **Senyurek, V.**, **Kurum, M.**, **Gurbuz, A.C.**, **Boyd, D.**, **Moorhead, R.**, **Crow, W.T. and Eroglu, O.** (2022). Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations, *Remote Sensing of Environment*, *276*, 113041.

[78] **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 Sensing*, *7*, 16398–16421.

[79] **Yuan, Q.**, **Shen, H.**, **Li, T.**, **Li, Z.**, **Li, S.**, **Jiang, Y.**, **Xu, H.**, **Tan, W.**, **Yang, Q.**, **Wang, J.**, **Gao, J. and Zhang, L.** (2020). Deep learning in environmental remote sensing: achievements and challenges, *Remote Sensing of Environment*, *241*, 111716.

[80] **El Hajj, M.**, **Baghdadi, N.**, **Zribi, M. and Bazzi, H.** (2017). Synergic use of Sentinel-1 and Sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas, *Remote Sensing*, *9*(12), 1292.

[81] **Hegazi, E.H.**, **Yang, L. and Huang, J.** (2021). A convolutional neural network algorithm for soil moisture prediction from Sentinel-1 SAR images, *Remote Sensing*, *13*(24), 4964.

[82] **Chung, J.**, **Lee, Y.**, **Kim, J.**, **Jung, C. and Kim, S.** (2022). Soil moisture content estimation based on Sentinel-1 SAR imagery using an artificial neural network and hydrological components, *Remote Sensing*, *14*(3), 465.

[83] **Chaudhary, S.K.**, **Srivastava, P.K.**, **Gupta, D.K.**, **Kumar, P.**, **Prasad, R.**, **Pandey, D.K.**, **Das, A.K. and Gupta, M.** (2022). Machine learning algorithms for soil moisture estimation using Sentinel-1: model development and implementation, *Advances in Space Research*, *69*(4), 1799–1812.

[84] **Cui, H.**, **Jiang, L.**, **Paloscia, S.**, **Santi, E.**, **Pettinato, S.**, **Wang, J.**, **Fang, X. and Liao, W.** (2022). The potential of ALOS-2 and Sentinel-1 radar data for soil moisture retrieval with high spatial resolution over agroforestry areas, China, *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–17.

[85] **Nativel, S.**, **Ayari, E.**, **Rodriguez-Fernandez, N.**, **Baghdadi, N.**, **Madelon, R.**, **Albergel, C. and Zribi, M.** (2022). Hybrid methodology using Sentinel-1/Sentinel-2 for soil moisture estimation, *Remote Sensing*, *14*(10), 2434.

[86] **Eroglu, O.**, **Kurum, M.**, **Boyd, D. and Gurbuz, A.C.** (2019). High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks, *Remote Sensing*, *11*, 2272.

[87] **Hachani, A.**, **Ouessar, M.**, **Paloscia, S.**, **Santi, E. and Pettinato, S.** (2019). Soil moisture retrieval from Sentinel-1 acquisitions in an arid environment in Tunisia: application of artificial neural networks techniques, *International Journal of Remote Sensing*, *40*, 9159–9180.

[88] **Lee, C.S.**, **Sohn, E.**, **Park, J.D. and Jang, J.D.** (2019). Estimation of soil moisture using deep learning based on satellite data: a case study of South Korea, *GIScience & Remote Sensing*, *56*, 43–67.

[89] **Pascanu, R.**, **Mikolov, T. and Bengio, Y.** (2013). On the difficulty of training recurrent neural networks, *International conference on machine learning*, PMLR, pp.1310–1318.

[90] **Hochreiter, S. and Schmidhuber, J.** (1997). Long short-term memory, *Neural computation*, *9*(8), 1735–1780.

[91] **Fang, K.**, **Shen, C.**, **Kifer, D. and Yang, X.** (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental US using a deep learning neural network, *Geophysical Research Letters*, *44*(21), 11–030.

[92] **Fang, K.**, **Pan, M. and Shen, C.** (2018). The value of SMAP for long-term soil moisture estimation with the help of deep learning, *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233.

[93] **Fang, K. and Shen, C.** (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel, *Journal of Hydrometeorology*, *21*, 399–413.

[94] **Ming, W.**, **Ji, X.**, **Zhang, M.**, **Li, Y.**, **Liu, C.**, **Wang, Y. and Li, J.** (2022). A hybrid triple collocation-deep learning approach for improving soil moisture estimation from satellite and model-based data, *Remote Sensing*, *14*, 1744.

[95] **Dorigo, W.**, **Wagner, W.**, **Hohensinn, R.**, **Hahn, S.**, **Paulik, C.**, **Xaver, A.**, **Gruber, A.**, **Drusch, M.**, **Mecklenburg, S.**, **van Oevelen, P.** *et al.* (2011). The international soil moisture network: a data hosting facility for global in situ soil moisture measurements, *Hydrology and Earth System Sciences*, *15*(5), 1675–1698.

[96] **Dorigo, W.**, **Himmelbauer, I.**, **Aberer, D.**, **Schremmer, L.**, **Petrakovic, I. and Zappa, L. ... Sabia, R.** (2021). The international soil moisture network: serving Earth system science for over a decade, *Hydrology and Earth System Sciences*, *25*(11), 5749–5804, https://hess.copernicus.org/articles/25/5749/2021/.

[97] **Montzka, C.**, **Bogena, H.R.**, **Herbst, M.**, **Cosh, M.H.**, **Jagdhuber, T. and Vereecken, H.** (2021). Estimating the number of reference sites necessary for the validation of global soil moisture products, *IEEE Geoscience and Remote Sensing Letters*, *18*(9), 1530–1534.

[98] **European Space Agency** (2017). Land cover CCI product user guide version 2 technical report, *Available at: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, (accessed on 18 July 2022).*

[99] **Rubel, F.**, **Brugger, K.**, **Haslinger, K. and Auer, I.** (2017). The climate of the European Alps: shift of very high resolution Köppen-Geiger climate zones 1800–2100, *Meteorologische Zeitschrift*, *26*(2), 115–125, http://www.schweizerbart.de/papers/metz/detail/26/87237/The_climate_of_the_European_Alps_Shift_of_very_hig?af=crossref.

[100] **Gorelick, N.**, **Hancher, M.**, **Dixon, M.**, **Ilyushchenko, S.**, **Thau, D. and Moore, R.** (2017). Google earth engine: planetary-scale geospatial analysis for everyone, *Remote sensing of Environment*, *202*, 18–27.

[101] **Liu, Y.**, **Qian, J. and Yue, H.** (2021). Combined Sentinel-1A with Sentinel-2A to estimate soil moisture in farmland, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 1292–1310.

[102] **Baghdadi, N.N.**, **El Hajj, M.**, **Zribi, M. and Fayad, I.** (2016). Coupling SAR C-Band and optical data for soil moisture and leaf area index retrieval over irrigated grasslands, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(3), 1229–1243.

[103] **Bazzi, H.**, **Baghdadi, N.**, **El Hajj, M.**, **Zribi, M. and Belhouchette, H.** (2019). A comparison of two soil moisture products S^2MP and Copernicus-SSM over Southern France, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(9), 3366–3375.

[104] **Liang, J.**, **Liang, G.**, **Zhao, Y. and Zhang, Y.** (2021). A synergic method of Sentinel-1 and Sentinel-2 images for retrieving soil moisture content in agricultural regions, *Computers and Electronics in Agriculture*, *190*, 106485.

[105] **Palmisano, D.**, **Satalino, G.**, **Balenzano, A. and Mattia, F.** (2022). Coherent and incoherent change detection for soil moisture retrieval from Sentinel-1 data, *IEEE Geoscience and Remote Sensing Letters*, 1–1.

[106] **Entekhabi, D.**, **Yueh, S.**, **O'Neill, P.**, **Kellogg, K.**, **Allen, A.**, **Bindlish, R.**, **Brown, M.**, **Chan, S.**, **Colliander, A.**, **Crow, W.** *et al.* (2014). SMAP handbook soil moisture active passive, *Mapping Soil Moisture Freeze/Thaw from Space, Pasadena, CA*.

[107] **Takaku, J.**, **Tadono, T. and Tsutsui, K.** (2014). Generation of high-resolution global DSM from ALOS Prism, *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *2*(4).

[108] **Longden, A.J.** Meteomatics, *102nd American Meteorological Society Annual Meeting*, AMS.

[109] **Goodfellow, I.**, **Bengio, Y. and Courville, A.** (2016). *Deep learning*, MIT press.

[110] **Sun, R.Y.** (2020). Optimization for deep learning: an overview, *Journal of the Operations Research Society of China*, *8*, 249–294.

[111] **Bai, J.**, **Cui, Q.**, **Zhang, W. and Meng, L.** (2019). An approach for downscaling SMAP soil moisture by combining Sentinel-1 SAR and MODIS data, *Remote Sensing*, *11*(23), 2736.

[112] **Millard, K. and Richardson, M.** (2018). Quantifying the relative contributions of vegetation and soil moisture conditions to polarimetric C-Band SAR response in a temperate peatland, *Remote Sensing of Environment*, *206*, 123–138.

[113] **Çelik, M.F. and Erten, E.** (2022). Biophysical parameter estimation of crops from polarimetric synthetic aperture radar imagery with data-driven polynomial chaos expansion and global sensitivity analysis, *Computers and Electronics in Agriculture*, *194*, 106781.

[114] **Benninga, H.J.F.**, **van der Velde, R. and Su, Z.** (2019). Impacts of radiometric uncertainty and weather-related surface conditions on soil moisture retrievals with Sentinel-1, *Remote sensing*, *11*(17), 2025.

[115] **Yuzugullu, O.**, **Lorenz, F.**, **Fröhlich, P. and Liebisch, F.** (2020). Understanding fields by remote sensing: soil zoning and property mapping, *Remote Sensing*, *12*(7), https://www.mdpi.com/2072-4292/12/7/1116.

[116] **Mateo-Sanchis, A.**, **Piles, M.**, **Muñoz-Marí, J.**, **Adsuara, J.E.**, **Pérez-Suay, A. and Camps-Valls, G.** (2019). Synergistic integration of optical and microwave satellite data for crop yield estimation, *Remote Sensing of Environment*, *234*, 111460.

[117] **Tian, H.**, **Wang, P.**, **Tansey, K.**, **Han, D.**, **Zhang, J.**, **Zhang, S. and Li, H.** (2021). A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China, *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102375.

[118] **Kang, X.**, **Huang, C.**, **Zhang, L.**, **Zhang, Z. and Lv, X.** (2022). Downscaling solar-induced chlorophyll fluorescence for field-scale cotton yield estimation by a two-step convolutional neural network, *Computers and Electronics in Agriculture*, *201*, 107260.

[119] **Kamir, E.**, **Waldner, F. and Hochman, Z.** (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods, *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 124–135.

[120] **Sawan, Z.M.** (2017). Cotton production and climatic factors: Studying the nature of its relationship by different statistical methods, *Cogent Biology*, *3*(1), 1292882.

[121] **Celik, M.F.**, **Isik, M.S.**, **Yuzugullu, O.**, **Fajraoui, N. and Erten, E.** (2022). Soil moisture prediction from remote sensing images coupled with climate, soil texture and topography via deep learning, *Remote Sensing*, *14*(21).

[122] **Guo, W.**, **Maas, S.J. and Bronson, K.F.** (2012). Relationship between cotton yield and soil electrical conductivity, topography, and Landsat imagery, *Precision Agriculture*, *13*(6), 678–692.

[123] **Yang, Y.**, **Wu, J.**, **Du, Y.L.**, **Gao, C.**, **Tang, D.W.S. and van der Ploeg, M.** (2022). Effect on soil properties and crop yields to long-term application of superabsorbent polymer and manure, *Frontiers in Environmental Science*, *10*.

[124] **Kang, Y.**, **Ozdogan, M.**, **Zhu, X.**, **Ye, Z.**, **Hain, C. and Anderson, M.** (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest, *Environmental Research Letters*, *15*(6), 064005.

[125] **Ma, Y.**, **Zhang, Z.**, **Kang, Y. and Özdoğan, M.** (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a bayesian neural network approach, *Remote Sensing of Environment*, *259*, 112408.

[126] **Gómez, D.**, **Salvador, P.**, **Sanz, J. and Casanova, J.L.** (2021). Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico, *Agricultural and Forest Meteorology*, *300*, 108317.

[127] **Li, Z.**, **Ding, L. and Xu, D.** (2022). Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China, *Science of The Total Environment*, *815*, 152880.

[128] **Mateo-Sanchis, A.**, **Adsuara, J.E.**, **Piles, M.**, **Munoz-Marí, J.**, **Perez-Suay, A. and Camps-Valls, G.** (2023). Interpretable long short-term memory networks for crop yield estimation, *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5.

[129] **Kaur Dhaliwal, J.**, **Panday, D.**, **Saha, D.**, **Lee, J.**, **Jagadamma, S.**, **Schaeffer, S. and Mengistu, A.** (2022). Predicting and interpreting cotton yield and its determinants under long-term conservation management practices using machine learning, *Computers and Electronics in Agriculture*, *199*, 107107.

[130] **Wolanin, A.**, **Mateo-García, G.**, **Camps-Valls, G.**, **Gómez-Chova, L.**, **Meroni, M.**, **Duveiller, G.**, **Liangzhi, Y. and Guanter, L.** (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt, *Environmental Research Letters*, *15*(2), 024019.

[131] **Nori, H.**, **Jenkins, S.**, **Koch, P. and Caruana, R.** (2019). Interpretml: a unified framework for machine learning interpretability, *arXiv preprint arXiv:1909.09223*.

[132] **Maxwell, A.E. and Shobe, C.M.** (2022). Land-surface parameters for spatial predictive mapping and modeling, *Earth-Science Reviews*, *226*, 103944.

[133] **Deger, Z.T.**, **Kaya, G.T. and Wallace, J.W.** (2023). *Estimate deformation capacity of non-ductile RC shear walls using explainable boosting machine*, 2301.04652.

[134] **Smith K. Khare, U.R.A.** (2023). An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals, *Computers in Biology and Medicine*, *155*, 106676.

[135] **USDA**. *United States Department of Agriculture National Agricultural Statistics Service*, https://quickstats.nass.usda.gov/.

[136] **Johnson, D.M.** (2016). A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products, *International Journal of Applied Earth Observation and Geoinformation*, *52*, 65–81.

[137] **Konings, A.G.**, **Piles, M.**, **Das, N. and Entekhabi, D.** (2017). L-band vegetation optical depth and effective scattering albedo estimation from SMAP, *Remote Sensing of Environment*, *198*, 460–470.

[138] **Thornton, M.**, **Shrestha, R.**, **Wei, Y.**, **Thornton, P.**, **Kao, S.C. and Wilson, B.** (2022). *Daymet: daily surface weather data on a 1-km grid for North America, version 4*.

[139] **Poggio, L.**, **de Sousa, L.M.**, **Batjes, N.H.**, **Heuvelink, G.B.M.**, **Kempen, B.**, **Ribeiro, E. and Rossiter, D.** (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, *Soil*, *7*(1), 217–240.

[140] **Lou, Y.**, **Caruana, R.**, **Gehrke, J. and Hooker, G.** (2013). Accurate intelligible models with pairwise interactions, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, p.623–631.

[141] **Akiba, T.**, **Sano, S.**, **Yanase, T.**, **Ohta, T. and Koyama, M.** (2019). Optuna: a next-generation hyperparameter optimization framework, *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

CURRICULUM VITAE

Name Surname: Mehmet Furkan ÇELİK

EDUCATION ⋮

- **B.Sc.** ⋮ 2013, Istanbul Technical University, Faculty of Civil Engineering, Department of Geomatics Engineering
- **M.Sc.** ⋮ 2016, Istanbul Technical University, Institute of Science, Engineering, and Technology, Department of Geomatics Engineering

PUBLICATIONS AND PRESENTATIONS ON THE THESIS:

- **Çelik, M.F., Işık, M.S. Taşkın, G., Erten, E. & Camps-Valls, G.** (2023). *Explainable Artificial Intelligence for Cotton Yield Prediction with Multisource Data*, IEEE Geoscience and Remote Sensing Letters, (Accepted), https://doi.org/10.1109/LGRS.2023.3303643.
- **Çelik, M.F., Işık, M.S., Yüzügüllü, O., Fajraoui, N. & Erten, E.** (2022). *Soil Moisture Prediction from Remote Sensing Images Coupled with Climate, Soil Texture and Topography via Deep Learning*, Remote Sensing, 14(21), https://doi.org/10.3390/rs14215584.
- **Çelik, M.F. & Erten, E.** (2022). *Biophysical parameter estimation of crops from polarimetric synthetic aperture radar imagery with data-driven polynomial chaos expansion and global sensitivity analysis*, Computers and Electronics in Agriculture, 194(106781), https://doi.org/10.1016/j.compag.2022.106781.
- **Çelik, M.F., Işık, M.S., Erten, E., Camps-Valls, G.** (2023). *Explainability of end and mid-season cotton yield predictors in CONUS*, IEEE International Geoscience and Remote Sensing Symposium IGARSS, 16-21 July 2023, IEEE.
- **Çelik, M.F., Işık, M.S., Erten, E., Gulsen, T.** (2023). *Informative Earth Observation Variables for Cotton Yield Prediction Using Explainable Boosting Machine*, IEEE International Geoscience and Remote Sensing Symposium IGARSS, 16-21 July 2023, IEEE.

- **Celik, M. F., Isik, M. S., Yuzugullu, O., Fajraoui, F., Erten, E.** (2023). *LSTM ile Toprak Neminin Tahmini için Çok Kaynaklı Veri Entegrasyonu | Multi-source Data Fusion for Estimation of Soil Moisture with LSTM*, Türkiye Ulusal Fotogrametri ve Uzaktan Algılama Birliği (TUFUAB) XII. Teknik Sempozyumu, 24-26 Mayıs, Sivas, Türkiye.

- **Çelik, M.F. & Erten, E.** (2021). *Principal component analysis based polynomial chaos expansion regression of leaf area index from PolSAR imagery*, IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021, pp. 6096-6099, https://doi.org/10.1109/IGARSS47720.2021.9554929

OTHER PUBLICATIONS AND PRESENTATIONS:

- **Çelik, M.F., Işık, M.S., and Erten, E.** (2023). *Interpretable Cotton Yield Prediction Model Using Earth Observation Time Series*, IEEE International Geoscience and Remote Sensing Symposium IGARSS, 16-21 July 2023, pp. 3442-3445 .

- **Erten, E., Çelik, M.F., Şahin, Z.M.**, (2018), *TANDEM-X Sayısal Yükseklik Modelinin Oluşturulması.*, Harita Dergisi, 160, 47-54.

- **Yağcı, O., Yildirim, I., Çelik, M.F., Kitsikoudis, V., Kirca, V.S.O. & Duran, Z.**, (2017), *Clear Water scour around a finite array of cylinders.* Applied Ocean Research, 68, 114-129, https://doi.org/10.1016/j.coastaleng.2017.09.001

- **Kırca, V.S.O., Kitsikoudis, V., Yağcı, O. & Çelik, M.F.**, (2017), *Clear-water scour and flow field alteration around an inclined pile.* Coastal Engineering, 128, 59-73, https://doi.org/10.1016/j.advwatres.2016.10.002

- **Öztürk, O., Bilgilioğlu, B.B., Çelik, M.F., Bilgilioğlu, S.S., Uluğ, R.**, (2017), *İnsanız Hava Aracı (İHA) Görüntüleri İle Ortofoto Üretiminde Yükseklik Ve Kamera Açısının Doğruluğa Etkisinin Araştırılması.*, Geomatik Dergisi, 3, 135-142, https://doi.org/10.29128/geomatik.327049.

- **Erten, E., Cristian, R., Juanma, L.S., Celik, M.F.**, (2017), *Interferometric Sar For Characterization Of Wetland Lakes As A Function Of Suspending Sediment Cover And Depth.* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 6239-6242, https://doi.org/10.1109/IGARSS.2017.8128435.

- **Düşgün, C., Şahin, Z.M., Görgülüoğlu, B., Celik, M.F., Bilgilioğlu, B.B., Doğru, A.Ö., Erten, E.**, (2017), *3D Modelling of ITU Ayazaga Campus.* Internatıonal Symposium On GIS Applications In Geography Geosciences.

- **Yalçın, H., Celik, M.F., Erten, E.**, (2017), *Detecting rooftops in 3D point clouds for solar mapping.* 25th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, https://doi.org/10.1109/SIU.2017.7960493.

- **Yalçın, H., Celik, M.F.**, (2017), *Neighbourhood characterization using visual architectural features.*, 25th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, https://doi.org/10.1109/SIU.2017.7960732.

- **Avsar, E.O., Celik, M.F., Bindir, E., Arslan, A.E., Cokkececi, D., Seker, D.Z., Pala, S.**, (2016), *Deformation monitoring of retrofitted short concrete colums with laser sensor.* ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 765-769, https://doi.org/10.5194/isprs-archives-XLI-B5-765-2016.

- **Yağcı, O., Çelik, M.F., Kitsikoudis, V., Kirca, V.S.O., Hodoglu, C., Valyrakis, M., Duran, Z. & Kaya, S.**, (2016), *Scour patterns around isolated vegetation elements.* Advances in Water Resources, 97, 251-265, https://doi.org/10.1016/j.advwatres.2016.10.002

- **Yagci, O., Kitsikoudis, V., Celik, M.F., Hodoglu, C., Kirca, V.S.O., Valyrakis, M., Duran, Z., Kaya, S.**, (2015), *The variation of local scour pattern around representative natural vegetation elements.* 36th IAHR Worl Congress.

- **Cakir, Z., Aslan, G., Dogan, U., Kaya, S., Ergintav, S., Oz, D., Celik, M.F.**, (2015), *Surface creep along the 1999 Izmit earthquake's rupture (Turkey) from InSAR, GPS and terrestrial LIDAR.* AGU Fall Meeting Abstracts, pp. G21A-1007

- **Denli, H., Celik, M.F., Kaya, S., Duran, Z.**, (2014), *Investigation of the Objects Depending on Distance Scanned with Laser Scanner*. AGU Fall Meeting Abstracts, pp. G31A-0393,