DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PREDICTION AND FORECASTING OF DROUGHT AREAS BASED ON REMOTE SENSING DATA USING MACHINE LEARNING METHODS

by
Gökçe GÖK

February, 2025
İZMİR

PREDICTION AND FORECASTING OF DROUGHT AREAS BASED ON REMOTE SENSING DATA USING MACHINE LEARNING METHODS

A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science in Mathematics

by
Gökçe GÖK

February, 2025
İZMİR

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**PREDICTION AND FORECASTING OF DROUGHT AREAS BASED ON REMOTE SENSING DATA USING MACHINE LEARNING METHODS**" completed by **GÖKÇE GÖK** under supervision of **PROF. DR. BURCU HÜDAVERDİ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Prof. Dr. Burcu HÜDAVERDİ

Yönetici

.....	.....
Prof. Dr. Neslihan DEMİREL	Prof. Dr. Gözde ULUTAGAY
Jury Member	Jury Member

Prof. Dr. Abdullah SEÇGİN
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor, Prof. Dr. Burcu HÜDAVERDİ, for her guidance and support throughout my thesis. This study would not have been possible without her patience, encouragement, and the delightful humor that lightened even the most challenging moments.

I also extend my sincere thanks to Prof. Dr. Ergin TARI. His belief in me, along with his curiosity and discipline, has always continued to inspire and motivate me.

I am grateful to my parents for their support in every possible way. I also appreciate everyone who has supported me — whether through their kind words, helpful actions, or simply by being there.

Finally, I acknowledge the Scientific and Technological Research Council of Turkey (TÜBİTAK) for supporting project 1002-A under project number 123F362.

Gökçe GÖK

PREDICTION AND FORECASTING OF DROUGHT AREAS BASED ON REMOTE SENSING DATA USING MACHINE LEARNING METHODS

ABSTRACT

The reliance on historical meteorological data for drought forecasting faces increasing challenges due to the impact of climate change. The complex and interconnected nature of climate systems further amplifies the difficulty in accurately forecasting droughts. Remote sensing data and machine learning methods have frequently been used in the prediction and forecasting of drought areas, due to their high accuracy and efficiency.

In the first part of this study, a comparative analysis has been conducted using Random Forests, Gradient Boosting Machines, eXtreme Gradient Boosting, and Light Gradient Boosting Machines methods to predict short-term droughts. Landsat 8 images and SRTM DEM 30 m data have been utilized to predict Soil Moisture Index (SMI) values for the severe drought areas. Finally, the importance of features in predicting drought events has been determined, highlighting the aspects that most critically affect drought occurrence.

In the second part, Landsat 8 images have been collected sequentially for SMI forecasting. It has been performed on two versions of the agricultural area dataset to examine the effect of cloud cover on forecasting, using Convolutional Long Short-Term Memory (ConvLSTM). Since the aim is to forecast future drought severity, a multi-step prediction approach has been employed through the ConvLSTM processes.

Through this two-part approach, the study offers a detailed examination of machine learning methods for SMI prediction in severe drought areas and explores the potential of ConvLSTM in forecasting drought impacts in agricultural contexts.

Keywords: Drought forecasting, soil moisture index, remote sensing, machine learning, ConvLSTM

UZAKTAN ALGILAMA VERİLERİNE DAYALI KURAKLIK ALANLARININ MAKİNE ÖĞRENME YÖNTEMLERİ KULLANILARAK TAHMİNİ VE ÖNGÖRÜSÜ

ÖZ

Kuraklık öngörümlemede, geçmiş meteorolojik verilere olan bağımlılık iklim değişikliğinin etkisi nedeniyle artan zorluklara sebep olmaktadır. İklim sistemlerinin karmaşık ve birbirine bağlı doğası, kuraklığı doğru bir şekilde tahmin etme zorluğunu daha da artırmaktadır. Uzaktan algılama verileri ve makine öğrenimi yöntemleri, yüksek doğruluk ve verimlilikleri sayesinde kuraklık bölgelerinin tahmini ve öngörümlenmesinde sıklıkla kullanılmaktadır.

Bu çalışmanın ilk bölümünde, kısa vadeli kuraklıkları tahmin etmek için Rastgele Ormanlar (Random Forests), Gradyan Artırma Makineleri (Gradient Boosting Machines), Aşırı Gradyan Artırma (eXtreme Gradient Boosting) ve Hafif Gradyan Artırma Makineleri (Light Gradient Boosting Machines) yöntemleri kullanılarak karşılaştırmalı bir analiz yapılmıştır. Şiddetli kuraklık alanları için Toprak Nem İndeksi (Soil Moisture Index - SMI) değerlerini tahmin etmek amacıyla Landsat 8 görüntüleri ve SRTM DEM (30 m) verileri kullanılmıştır. Son olarak, kuraklık olaylarını tahmin etmede özelliklerin önemi belirlenmiş ve kuraklık oluşumunu en kritik şekilde etkileyen unsurlar vurgulanmıştır.

İkinci bölümde, SMI öngörümlenmesi için Landsat 8 görüntüleri ardışık olarak toplanmıştır. Öngörümleme işlemi, bulut örtüsünün tahmin üzerindeki etkisini incelemek amacıyla, tarımsal alan veri setinin iki farklı versiyonu üzerinde, Evrişimsel Uzun Kısa Süreli Bellek (Convolutional Long Short-Term Memory - ConvLSTM) kullanılarak gerçekleştirilmiştir. Gelecekteki kuraklığın şiddetini tahmin etmek amaçlandığından, ConvLSTM süreçleri aracılığıyla çok adımlı bir tahmin yaklaşımı kullanılmıştır.

Bu iki aşamalı yaklaşım sayesinde, çalışma şiddetli kuraklık bölgelerinde SMI

tahmini için makine öğrenimi yöntemlerini detaylı bir şekilde incelemekte ve tarımsal bağlamlarda kuraklık etkilerinin öngörümlenmesinde ConvLSTM'nin potansiyelini araştırmaktadır.

Anahtar kelimeler: Kuraklık tahmini, toprak nemlilik endeksi, uzaktan algılama, makine öğrenmesi, ConvLSTM

CONTENTS

Page

M.Sc. THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
LIST OF FIGURES ix
LIST OF TABLES xii
LIST OF ABBREVIATIONS xiii

CHAPTER ONE – INTRODUCTION 1

CHAPTER TWO – METHODOLOGY 6

2.1 Land Surface Factors 6
2.1.1 Normalized Difference Vegetation Index 6
2.1.2 Enhanced Vegetation Index 7
2.1.3 Soil-Adjusted Vegetation Index 8
2.1.4 Modified Soil-Adjusted Vegetation Index 9
2.1.5 Slope and Aspect 11
2.1.6 Topographic Wetness Index 11
2.1.7 Normalized Difference Moisture Index 13
2.1.8 Modification of Normalized Difference Water Index 14
2.1.9 Moisture Stress Index 15
2.2 SMI Estimation 16
2.2.1 Land Surface Temperature 16
2.2.1.1 The NDVI threshold method 18
2.2.2 Soil Moisture Index 19
2.3 Machine Learning Models 20
2.3.1 Random Forests 21

2.3.2 Gradient Boosting Machines 22
2.3.3 XGBoost 24
2.3.4 LightGBM 25
2.3.5 ConvLSTM 27
2.4 Evaluation Metrics 29
2.4.1 R-squared 29
2.4.2 Mean Squared Error 30
2.4.3 Mean Absolute Error 30

CHAPTER THREE – DATASETS AND APPLICATION 32

3.1 Datasets 32
3.1.1 Drought Area Dataset 34
3.1.2 Agricultural Area Dataset 37
3.2 Application 39
3.2.1 SMI Prediction in Severe Drought Areas 39
3.2.1.1 Results using Random Forests 40
3.2.1.2 Results using GBM 42
3.2.1.3 Results using XGBoost 44
3.2.1.4 Results using LightGBM 46
3.2.2 SMI Forecasting in Agricultural Areas 50
3.2.2.1 Results using Raw Data 51
3.2.2.2 Results using Lag-Filled Data 55

CHAPTER FOUR – CONCLUSION 59

REFERENCES 63

LIST OF FIGURES

Page

Figure 2.1 Juxtaposition of the natural-color composite image (left) and the corresponding NDVI map (right) 7

Figure 2.2 Juxtaposition of the natural-color composite image (left) and the corresponding EVI map (right) 8

Figure 2.3 Juxtaposition of the natural-color composite image (left) and the corresponding SAVI map (right) 9

Figure 2.4 Juxtaposition of the natural-color composite image (left) and the corresponding MSAVI map (right) 10

Figure 2.5 Juxtaposition of the natural-color composite image (left), the slope map (middle), and the aspect map (right). 11

Figure 2.6 Identification of flow directions using the D8 algorithm: (a) DEM grid values, (b) Flow direction codes, (c) DEM coded according to flow directions, (d) Symbolic representation of flow directions (Source: Öztürk & Kılıç (2013)) 13

Figure 2.7 Juxtaposition of the natural-color composite image (left) and the corresponding TWI map (right) 13

Figure 2.8 Juxtaposition of the natural-color composite image (left) and the corresponding NDMI map (right) 14

Figure 2.9 Juxtaposition of the natural-color composite image (left) and the corresponding MNDWI map (right) 15

Figure 2.10 Juxtaposition of the natural-color composite image (left) and the corresponding MSI map (right) 16

Figure 2.11 The Ts-NDVI space (Source: Zeng et al. (2004)) 19

Figure 2.12 Juxtaposition of the natural-color composite image, NDVI, LST and SMI 20

Figure 3.1 Bands of Landsat 8 images with rows corresponding to red, NIR, and TIRS-1 34

Figure 3.2 SRTM DEM data for the study area used in the SMI prediction process. 35

Figure 3.3 The study area of SMI prediction 36

Figure 3.4 A sample of sequential data with rows corresponding to red, NIR, and TIRS-1 bands 37
Figure 3.5 Samples from versions of agricultural area dataset 38
Figure 3.6 Comparison of Random Forests metrics for training and test data..... 40
Figure 3.7 Visualization of feature importance based on Random Forests 41
Figure 3.8 Visual representation of SMI prediction results using Random Forests .. 41
Figure 3.9 Comparison of GBM metrics for training and test data 42
Figure 3.10 Visualization of feature importance based on GBM 43
Figure 3.11 Visual representation of SMI prediction results using GBM 43
Figure 3.12 Comparison of XGBoost metrics for training and test data..... 44
Figure 3.13 Visualization of feature importance based on XGBoost 45
Figure 3.14 Visual representation of SMI prediction results using XGBoost 45
Figure 3.15 Comparison of LightGBM metrics for training and test data 46
Figure 3.16 Visualization of feature importance based on LightGBM 47
Figure 3.17 Visual representation of SMI prediction results using LightGBM 47
Figure 3.18 Comparison of R^2 metrics across the models 48
Figure 3.19 Comparison of RMSE metrics across the models 48
Figure 3.20 Comparison of visual outputs from SMI prediction models..... 48
Figure 3.21 SMI prediction results using Random Forests for Lake Tuz, Türkiye..... 49
Figure 3.22 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloudy target) 52
Figure 3.23 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloudy target)..... 52
Figure 3.24 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloud-free target) 52
Figure 3.25 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloud-free target) 53

Figure 3.26 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloudy target) 56
Figure 3.27 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloudy target) 56
Figure 3.28 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloud-free target) 57
Figure 3.29 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloud-free target) 57
Figure 3.30 Comparison of forecasting results using raw data and lag-filled data for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset 58

LIST OF TABLES

Page

Table 2.1 Emissivity values based on NDVI ranges 19
Table 3.1 Bands of the Landsat 8 satellite (Source: USGS (2019))..... 32
Table 3.2 Land surface factors, input variables and corresponding data sources (Adapted from Park et al. (2019)) 36
Table 3.3 Hyperparameters configured for training process 50
Table 3.4 ConvLSTM model hyperparameters..... 50
Table 3.5 Performance metrics for train, validation, and test data using raw data 51
Table 3.6 MSE values over time steps using raw and lag-filled data..... 54
Table 3.7 Performance metrics for train, validation, and test data using lag-filled data55

LIST OF ABBREVIATIONS

GRACE	: Gravity Recovery and Climate Experiment
ARIMA	: Autoregressive Integrated Moving Average
VTCI	: Vegetation Temperature Condition Index
SPI	: Standardized Precipitation Index
SPEI	: Standardized Precipitation Evapotranspiration Index
ERT	: Extremely Randomized Trees
DT	: Decision Trees
RF	: Random Forests
SR	: Solar Radiation
TR	: Thermal Radiation
CatBoost	: Categorical Boosting
XGBoost	: eXtreme Gradient Boosting
LightGBM	: Light Gradient Boosting Machine
SMI	: Soil Moisture Index
RMSE	: Root Mean Square Error
MAE	: Mean Absolute Error
R^2	: Coefficient of Determination
SAR	: Synthetic Aperture RADAR
OLI	: Operational Land Imager
MS	: Measured Soil Moisture
LST	: Land Surface Temperature
NDVI	: Normalized Difference Vegetation Index
GBM	: Gradient Boosting Machines
ConvLSTM	: Convolutional Long Short-Term Memory
EVI	: Enhanced Vegetation Index
SAVI	: Soil-Adjusted Vegetation Index
MSAVI	: Modified Soil-Adjusted Vegetation Index
TWI	: Topographic Wetness Index
NDMI	: Normalized Difference Moisture Index

MNDWI	: Modification of Normalized Difference Water Index
MSI	: Moisture Stress Index
NIR	: Near Infrared
SWIR-1	: Short-Wavelength Infrared-1
SWIR-2	: Short-Wavelength Infrared-2
TIRS-1	: Thermal Infrared-1
TIRS-2	: Thermal Infrared-2
TVDI	: Temperature Vegetation Dryness Index
TVI	: Transformed Vegetation Index
PAC	: Post Atmospheric Correction Reflectance
TOA	: Top of Atmosphere Reflectance
DN	: Digital Number
VFC	: Vegetation Fractional Coverage
SRTM	: Shuttle Radar Topography Mission
DEM	: Digital Elevation Model
CDI	: Comprehensive Drought Index
NDWI	: Normalized Difference Water Index
WSVI	: Water Supply Vegetation Index
NDII	: Normalized Difference Infrared Index
SSM	: Surface Soil Moisture
ISMN	: International Soil Moisture Network
GPMI	: Precipitation State Index
TCI	: Temperature Condition Index
MLR	: Multiple Linear Regression
CNN	: Convolutional Neural Network
CWSI	: Crop Water Stress Index
TSS	: Total Sum of Squares
RSS	: Residual Sum of Squares
MSE	: Mean Square Error
TIRS	: Thermal Infrared Sensor
NASA	: National Aeronautics and Space Administration

USGS	: U.S. Geological Survey
WRS-2	: Worldwide Reference System-2
NGA	: National Geospatial-Intelligence Agency

CHAPTER ONE
INTRODUCTION

Drought is a complex and chaotic natural disaster, characterized by its slow onset and cumulative long-term adverse environmental, social, and economic impacts on a global scale. The inherent complexity of drought poses challenges for both its precise definition and accurate forecasting (Çamalan et al., 2017). Wilhite & Glantz (1985) argue that the lack of a universally recognized definition for drought arises from its varied impacts depending on the context. Through their analysis of more than 150 definitions, the study categorized drought into four distinct types: meteorological, agricultural, hydrologic, and socio-economic.

Meteorological drought is characterized by prolonged periods of insufficient rainfall, which is typically measured by its intensity and duration. This type of drought can lead to agricultural drought when reduced precipitation results in higher evaporation rates and insufficient soil moisture, ultimately affecting farming activities such as crop production and livestock management. Hydrological drought, on the other hand, occurs when a prolonged absence of precipitation significantly reduces surface water and groundwater reserves, impacting water availability for agricultural, industrial, and domestic use. Finally, socio-economic drought describes the broader social and economic consequences of drought, which often arise from imbalances in resource availability and demand. These impacts can include economic losses in agriculture, health-related challenges, and even population displacement in extreme cases (American Meteorological Society (AMS), 2025).

Türkiye's geographical location within the dry-summer subtropical macro-climate zone contributes to various climate-related challenges, both under current climate conditions and in light of potential future changes. These factors make the country particularly vulnerable to medium to high climate risks (Türkeş, 2020). Based on the Türkiye Desertification Risk Map from Karagöz et al. (2017), 12.7% of the country's land is identified as having low desertification risk, 53.2% as moderate risk, and 25.5% as high risk.

Drought poses significant risks to both natural and human systems. These risks emphasize the critical need to understand the dynamics of drought and to develop effective strategies for its prediction and mitigation. According to Wilhite et al. (2007), drought preparedness efforts can be improved through the implementation of drought management approaches by governments and the creation of early warning mechanisms.They also highlight the critical need to assess how effectively these measures minimize monetary, societal, and ecological impacts to guide planning efforts focused on effective drought mitigation.

Remote sensing satellites offer an efficient and affordable method for drought forecasting and prediction, thanks to their sophisticated sensor technology within this scope. Remote sensing data are scalable, enabling both local and global studies, while their high spatial, spectral, and temporal resolution plays a crucial role in providing accurate and actionable insights.

However, as highlighted by AghaKouchak et al. (2015), while remote sensing technologies present significant opportunities, they also face challenges such as data continuity issues, uncertainties, sensor changes, and community acceptance. Furthermore, such as the Gravity Recovery and Climate Experiment (GRACE), generate data that spans approximately a decade, potentially restricting their usability in long-term climate analyses. Nevertheless, these datasets still offer valuable insights into the hydrological and ecological processes linked to droughts.

Thanks to the comprehensive information, remotely-sensed data has been extensively applied in numerous researches related to drought monitoring, prediction, and forecasting. Recently, machine learning technologies have gained prominence in modeling and understanding drought characteristics. These algorithms are particularly effective for drought prediction as they can integrate large volumes of data derived from multiple sources, including satellite images, meteorological stations, and soil monitoring devices. By identifying patterns and relationships that may not be immediately apparent to humans, machine learning models can provide valuable insights. Moreover, they utilize past data to generate accurate forecasts

about future drought conditions, enabling better preparedness and mitigation of drought impacts on agriculture, water supplies, and ecosystems.

In the study conducted by Han et al. (2010) using remotely sensed data from the Guanzhong Plain, the autoregressive integrated moving average (ARIMA) models were used to forecast and simulate the Vegetation Temperature Condition Index (VTCI). The pixels corresponding to the locations of weather stations, which were averagely distributed across the study area, were utilized for analyzing the VTCI time series. In the AR(1) forecasting, the mean forecasting error was found to be 0.10 for a one-step forecast and 0.17 for a two-step forecast.

In the study by Rhee & Im (2017), South Korea was chosen as the study area for forecasting the Standardized Precipitation Index (SPI) and Standardized Precipitation Evapotranspiration Index (SPEI). This selection was based on its geographic location, which is particularly suitable for long-range climate forecasting. The SPI and SPEI indices are related to meteorological drought. Their comparative study demonstrated that machine learning methods provide better results than spatial interpolation techniques. Furthermore, the comparison of Extremely Randomized Trees (ERT), Decision Trees (DT), and Random Forests (RF) revealed that the drought prediction accuracy of ERT reaches up to 56% for SPI forecasts and 64% for SPEI forecasts when using long-range climate forecast data, outperforming the other methods.

In the study of Huang et al. (2023), input variables derived from the ERA5-Land dataset (Muñoz Sabater et al., 2021) were used for soil moisture prediction, including total precipitation (P), 10 m U-wind component (U), 10 m V-wind component (V), 2 m temperature (TA), surface net solar radiation (SR), and surface net thermal radiation (TR). The ConvLSTM model prediction was yielded a strong R^2 value of 0.92, demonstrating its high level of accuracy.

In the study conducted by Bojer et al. (2024), which examined the effects of drought on vegetation using meteorological drought indices and drought models, the Categorical Boosting (CatBoost) algorithm was identified as the most accurate

predictor for SPI. A comparison of the RF, eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost algorithms revealed that CatBoost achieved an overall accuracy of 94.25% and a kappa coefficient index of 0.9328.

The Soil Moisture Index (SMI) is one of the remote sensing indices for analyzing drought conditions. Soil moisture levels are a critical indicator for plant growth, agricultural production, and water resource management. The information on current water content of the soil provided by SMI makes it possible to detect drought conditions at an early stage.

In the study conducted by Park et al. (2019), the RF method was utilized for SMI prediction to enable short-term drought prediction without relying on meteorological data. Vegetation, thermal, water, and topographical land surface factors were used as input variables. For the training area, which is 20 times smaller than the study area in South Korea, the model obtained a root mean square error (RMSE) of 0.052, mean absolute error (MAE) of 0.039, and R^2 of 0.91. However, when applied to the entire training area, the evaluation metrics included an RMSE of 0.382, an MAE of 0.375, and an R^2 value of 0.58.

In the study conducted by Sutariya et al. (2021), Sentinel-1 C-band Synthetic Aperture RADAR (SAR) and Landsat-8 Operational Land Imager (OLI) data were utilized for soil moisture estimation. Through a multiple linear regression analysis performed using σ^0_{VV} polarization[1], gravimetrically measured soil moisture (MS%) from field observations, and the Soil Moisture Index (SMI), R^2 values of 0.788, 0.777, and 0.778 were achieved for different talukas in the study area.

This thesis aims to design and assess an SMI model for predicting severe drought conditions in Türkiye and to assess its applicability in forecasting agricultural droughts. The first part of the study focuses on modeling and predicting SMI in regions with significant drought risk using machine learning models, including RF,

[1]Backscattering coefficient for vertical-vertical polarization

Gradient Boosting Machines (GBM), XGBoost, and LightGBM. These predictions have been conducted for areas known to experience severe drought using indices such as the Normalized Difference Vegetation Index (NDVI) and Land Surface Temperature (LST) derived from Landsat 8 satellite images and SRTM DEM (30 m) data.

In the latter section of the research, a Convolutional Long Short-Term Memory (ConvLSTM) based model has been constructed to forecast drought conditions in agricultural areas. This process focuses on evaluating the ability of the SMI to anticipate drought impacts in agricultural contexts in Aydın, one of Türkiye's key agricultural regions. ConvLSTM layers are designed to manage spatio-temporal dependencies, making them highly suitable for analyzing sequential remote sensing data, e.g., for drought forecasting. Since the aim is to forecast future drought severity, a multi-step prediction approach has been employed through the ConvLSTM processes. Specifically, the model has been utilized 5 time steps as input to predict the 6th time step, with each time interval representing a 32-day period. The performance of the forecasting model was assessed through the Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics. This two-stage approach makes possible a detailed examination of the SMI's effectiveness in both severe drought areas and agricultural contexts.

The structure of this thesis is as follows: Chapter 2 provides an overview of the methodologies and techniques utilized in this study, along with a review of the relevant literature. Chapter 3 details the datasets used and describes the application processes and results for SMI prediction and forecasting. Finally, Chapter 4 concludes the study by outlining the results, exploring their significance, and suggesting directions for future studies.

CHAPTER TWO
METHODOLOGY

2.1 Land Surface Factors

Land surface factors provide crucial insights into the spatial and temporal patterns of drought. In this study, several indices obtained through remotely sensed data were utilized to predict the SMI. These land surface factors are categorized into four groups: vegetation, water, thermal, and topography. The vegetation factors include the NDVI, Enhanced Vegetation Index (EVI), Soil-Adjusted Vegetation Index (SAVI), and Modified Soil-Adjusted Vegetation Index (MSAVI). The topographic factors consist of slope, aspect, and the Topographic Wetness Index (TWI). The water factors include the Normalized Difference Moisture Index (NDMI), Modification of Normalized Difference Water Index (MNDWI), and Moisture Stress Index (MSI). Lastly, the thermal factors comprise the spectral bands of Landsat 8 imagery: Near Infrared (NIR), Short-Wavelength Infrared-1 (SWIR-1), Short-Wavelength Infrared-2 (SWIR-2), Thermal Infrared-1 (TIRS-1), and Thermal Infrared-2 (TIRS-2) (Park et al., 2019). The topographic, water, and vegetation factors are detailed below.

2.1.1 Normalized Difference Vegetation Index

The NDVI quantifies the state and concentration of vegetation by utilizing the reflectance values of the NIR and red bands of satellite imagery. The calculation formula is presented below as an Equation 2.1 (Tucker, 1979).

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{2.1}$$

This index spans values from $(-1, 1)$. Negative values refer a weakness or absence of water surfaces or vegetation, while positive values represent well-established and healthy plant cover.

In this study, NDVI has been primarily utilized for SMI estimation. The SMI values, have been estimated using the relationship between NDVI and LST as detailed in Section 2.2, were considered ground-truth values. Since the entire study is based on the prediction and forecasting of these SMI values, NDVI is one of the most critical parameters in this research. The juxtaposition of the natural-color composite image and the corresponding NDVI map is shown in Figure 2.1.

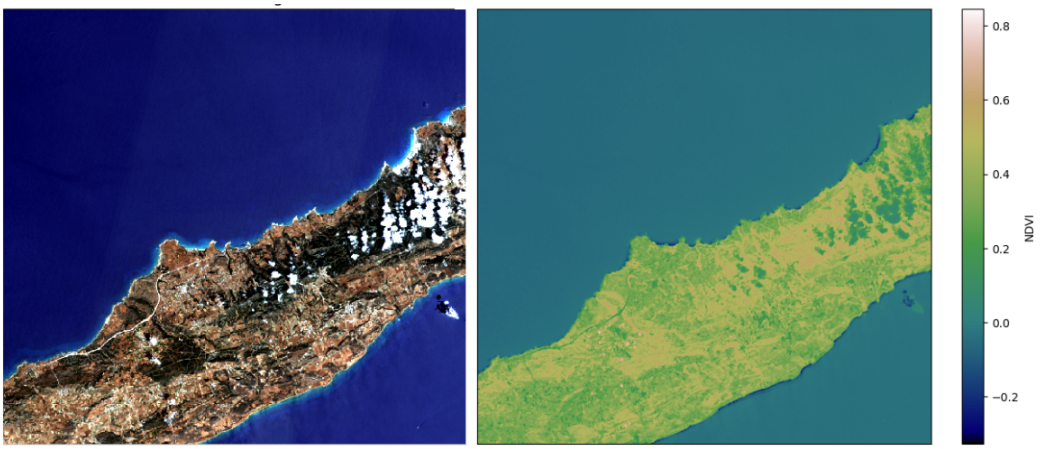


Figure 2.1 Juxtaposition of the natural-color composite image (left) and the corresponding NDVI map (right)

2.1.2 Enhanced Vegetation Index

The EVI can be used to quantify the health and density of vegetation cover on land surface similar to the NDVI. It is calculated using the blue bands of images in addition to the NIR and red bands. So, It gives extra information about vegetation. This index spans values from $(-1, 1)$, similar to NDVI.

As noted by USGS (2025a), EVI offers greater sensitivity compared to NDVI in regions with dense vegetation and corrects for atmospheric conditions as well as canopy background noise. The formula of EVI is presented below as an Equation 2.2.

$$EVI = G \cdot \frac{(NIR - Red)}{(NIR + C_1 * Red – C_2 * Blue + L)} \tag{2.2}$$

where, L value represents the canopy background adjustment parameter, while C

values denote the coefficients for atmospheric resistance. The coefficients specific to Landsat 8 are 2.5, 6, 7.5, and 1, corresponding to G, C_1, C_2, and L, respectively.

In their study, Liu et al. (2021) examined the effect of identifying the relationships between various vegetation indices and land surface temperature using different feature spaces on dryness monitoring. The study found that when NDVI did not reach saturation, Temperature Vegetation Dryness Index (TVDI) from bi-parabolic NDVI-Ts feature space demonstrated the highest effectiveness in dryness monitoring. Conversely, TVDI from EVI-Ts feature space was identified as the most suitable dryness index when NDVI approached or reached saturation.

In this study, EVI has been used as an input variable in the SMI prediction process. The juxtaposition of the natural-color composite image and the corresponding EVI map is shown in Figure 2.2.

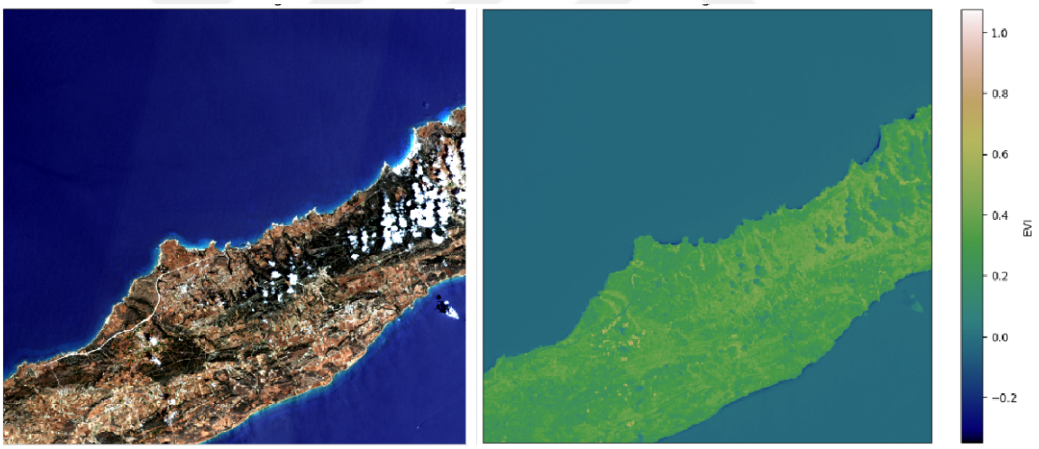


Figure 2.2 Juxtaposition of the natural-color composite image (left) and the corresponding EVI map (right)

2.1.3 Soil-Adjusted Vegetation Index

In the research conducted by Qi et al. (1994), The SAVI was introduced to reduce the impact of soil properties on the spectral data of vegetation canopies. It allows to evaluate soil effect along with vegetation health and density. Similar to NDVI, it is derived using the NIR and red bands of satellite images, with its value ranging between $(-1, 1)$. It adjusts the NDVI to account for the impact of soil brightness in regions

with sparse vegetation cover. The formula with the coefficients specific to Landsat 8 is presented below in Equation 2.3 (USGS, 2025a).

$$SAVI = 1.5 \cdot \frac{(NIR - Red)}{(NIR + Red + 0.5)} \tag{2.3}$$

In the study of Yıldırım & Asık (2018), NDVI and SAVI indices were utilized to assess agricultural drought. They observed a substantial correlation between NDVI/SAVI, temperature, and rainfall in agricultural regions, highlighting the impact of drought on reducing green vegetation density. It was further emphasized that the lowest NDVI/SAVI values occur under extreme drought conditions.

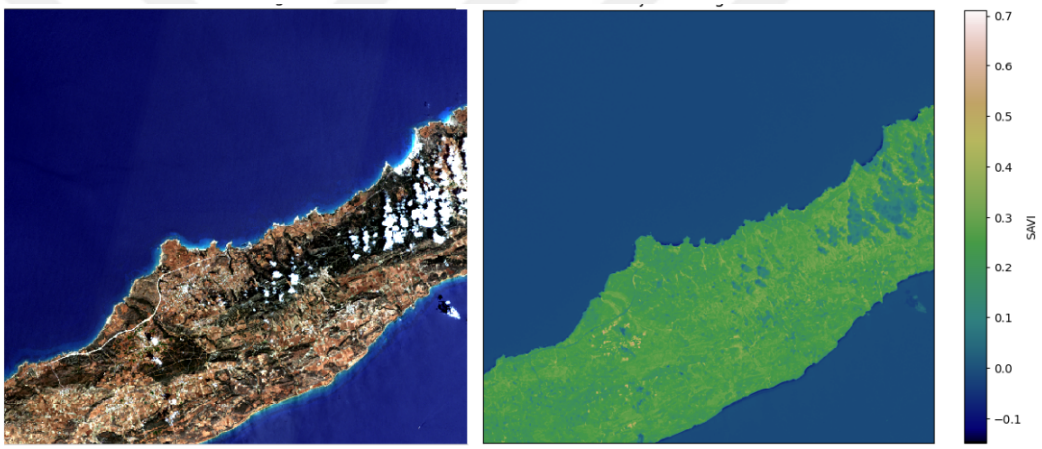


Figure 2.3 Juxtaposition of the natural-color composite image (left) and the corresponding SAVI map (right)

The juxtaposition of the natural-color composite image and the corresponding SAVI map is shown in Figure 2.3. In this study, the SAVI has been utilized as an input variable of SMI prediction.

2.1.4 Modified Soil-Adjusted Vegetation Index

The MSAVI is designed to further reduce the impact of bare soil on vegetation measurements, improving upon the SAVI. It is derived as a ratio of the red and NIR reflectance values, incorporating an inductive L function to enhance the suppression of soil effects on plant cover signals. The calculation formula is presented below as

an Equation 2.4 (USGS, 2025a).

$$MSAVI = \frac{2 \cdot NIR + 1 - \sqrt{(2 \cdot NIR + 1)^2 - 8 \cdot (NIR - Red)}}{2} \tag{2.4}$$

The last vegetation factor that has been used as an input variable for SMI prediction is MSAVI, and its juxtaposition with the natural-color composite image is shown in Figure 2.4.

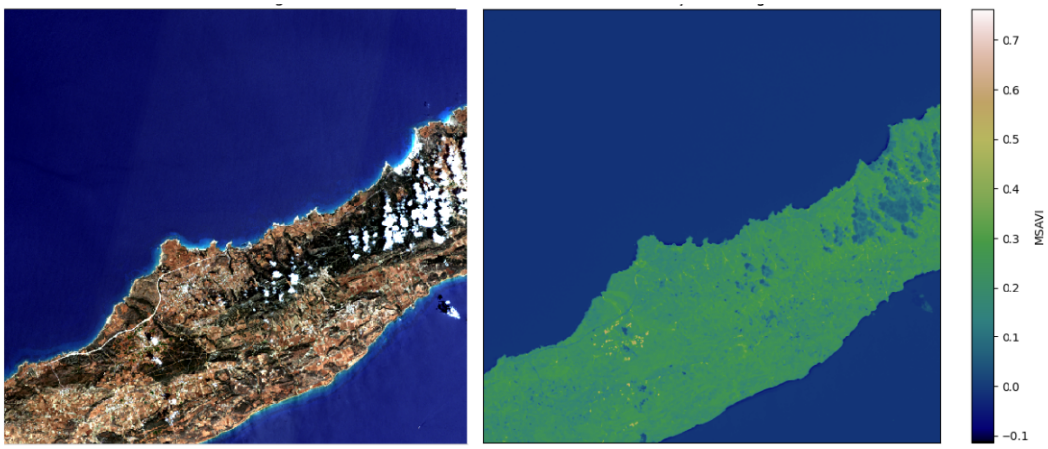


Figure 2.4 Juxtaposition of the natural-color composite image (left) and the corresponding MSAVI map (right)

In their study, Gu et al. (2009) derived NDVI, TVI (Transformed Vegetation Index), SAVI, and MSAVI indices using images processed with different radiometric correction methods, including post atmospheric correction reflectance (PAC), top of atmosphere reflectance (TOA), and digital number (DN). After examining the effects of various polynomial models, they observed that the accuracy of vegetation fractional coverage (VFC) estimations is influenced by radiometric corrections, vegetation indices, and model types. The highest R^2 value was achieved with cubic regression for all atmospheric correction methods. The highest R^2 and the lowest RMSE were achieved with the regression model establishing the relationship between NDVI and VFC.

2.1.5 Slope and Aspect

Slope and aspect are key topographic factors influencing drought conditions by affecting water retention, evaporation rates, and vegetation patterns, which in turn shape the soil moisture dynamics and drought vulnerability of a region. In this study, all topographic factors have been derived using Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) (30 m) data to investigate the influence of topography on SMI prediction. Specifically, the slope and aspect variables have been directly derived from the DEM. The juxtaposition of the natural-color composite image and the corresponding slope and aspect maps is shown in Figure 2.5.

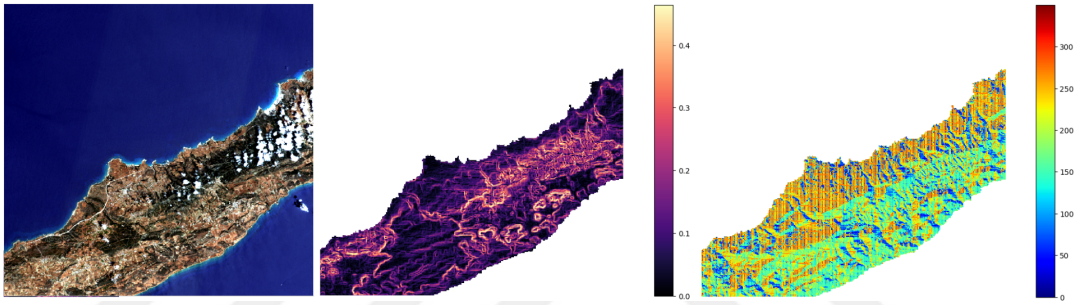


Figure 2.5 Juxtaposition of the natural-color composite image (left), the slope map (middle), and the aspect map (right).

Their study aimed at drought monitoring and modeling, Xu et al. (2024) calculated the Comprehensive Drought Index (CDI) for different ecological-geographic zones. Using a multivariate regression method, they utilized the DEM, slope, and aspect as terrain factors.

2.1.6 Topographic Wetness Index

TWI is an important measure used to determine the moisture potential of a region. TWI analyzes the influence of topography on water accumulation and flow direction and is derived based on the flow accumulation area and slope.

The calculation formula is presented below as an Equation 2.5 (Park et al., 2019).

$$TWI = \ln\left(\frac{a}{\tan b}\right) \tag{2.5}$$

Here, the variable a (m^2) represents the upstream contributing area, while b denotes the slope.

In their study, comparing the effectiveness of TWIs derived using different flow algorithms to predict soil moisture, Kopecký et al. (2021) stated that the specification of the flow algorithm is the most critical step influencing TWI performance. In this study, Deterministic single-flow (D8) algorithm (O'Callaghan & Mark, 1984) has been employed for TWI calculations.

In the D8 algorithm, flow within a cell is assumed to occur in the direction of the neighboring cell with the steepest slope and the lowest elevation. By considering a 3x3 window in a grid DEM, flow is directed to only one of the eight neighboring cells. Once the flow direction of each cell is determined, it becomes feasible to calculate the number of contributing cells for any given cell (Öztürk & Kılıç, 2013). Determination of flow directions using the D8 algorithm is shown in Figure 2.6.

The total number of cells coming from the upstream contributing area is calculated for each cell according to the determined flow directions with D8 algorithm. This total constitutes the upstream contributing area a value.

Within the scope of the D8 algorithm, the slope b is based on the elevation difference between each cell and its neighboring cell in the flow direction and is expressed by the following formula in Equation 2.6.

$$b = \arctan\left(\frac{\Delta z}{\Delta x}\right) \tag{2.6}$$

where Δx the elevation difference between the cell in the direction of flow and the current cell, while Δz represents the horizontal or diagonal distance depending on the

cell size.

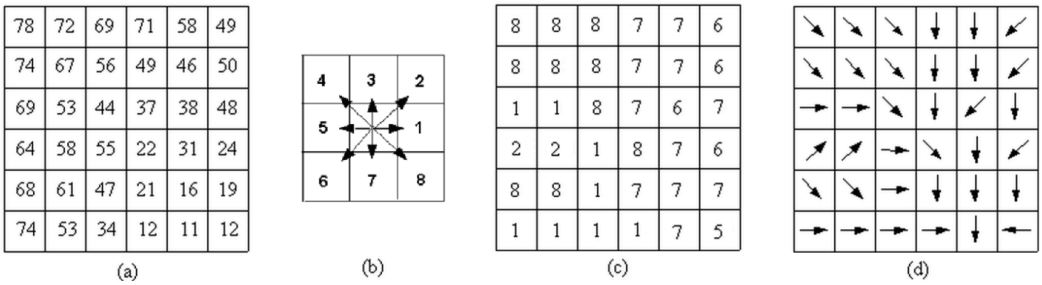


Figure 2.6 Identification of flow directions using the D8 algorithm: (a) DEM grid values, (b) Flow direction codes, (c) DEM coded according to flow directions, (d) Symbolic representation of flow directions (Source: Öztürk & Kılıç (2013))

The juxtaposition of the natural-color composite image and the corresponding TWI map is shown in Figure 2.7.

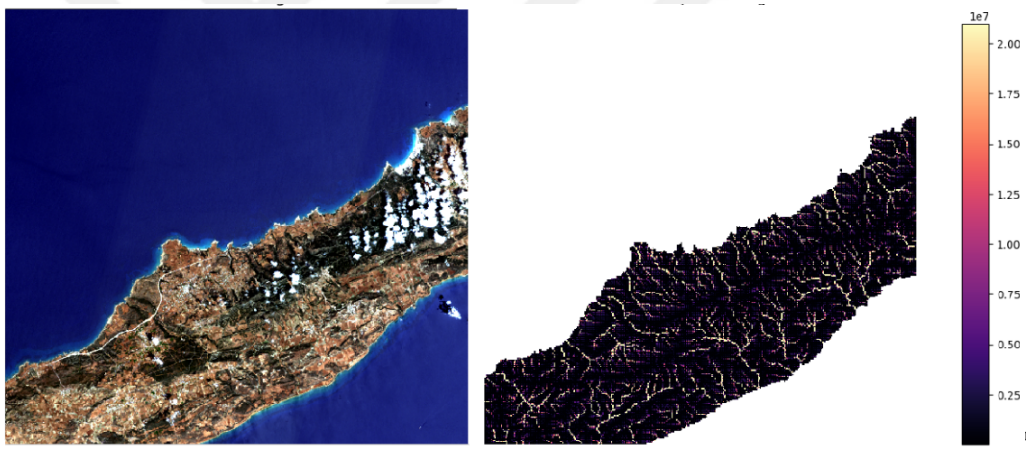


Figure 2.7 Juxtaposition of the natural-color composite image (left) and the corresponding TWI map (right)

2.1.7 Normalized Difference Moisture Index

NDMI is utilized to assess the moisture level in plant cover. It is computed using the NIR, and $SWIR_1$ bands of satellite images. The formula of NDMI is presented below as an Equation 2.7 (USGS, 2025a).

$$NDMI = \frac{(NIR - SWIR_1)}{(NIR + SWIR_1)} \tag{2.7}$$

In the study where the impact of multi-parameters on drought was analyzed, a high correlation between NDVI and NDMI was identified by Bhattacharya et al. (2021). It was emphasized that the use of multi-parameter indices is effective in drought monitoring due to the varying climatic conditions across regions.

In their study, NDVI, Normalized Difference Water Index (NDWI), NDMI, SAVI, and Land Surface Temperature (LST) were employed by Das et al. (2023) to investigate changes in land coverage. Using indices derived from Landsat data, they examined the changes in forests, water bodies, settlements, and agricultural areas.

The juxtaposition of the natural-color composite image and the corresponding NDMI map is shown in Figure 2.8.

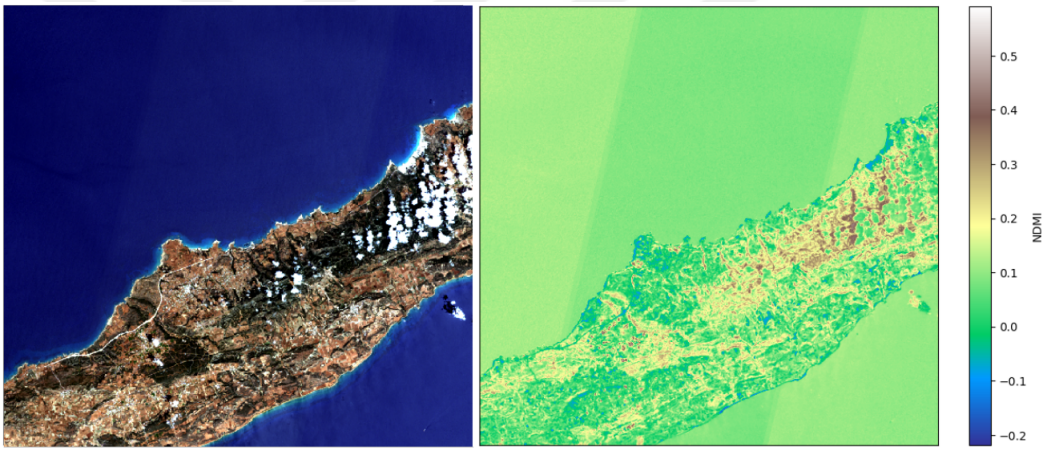


Figure 2.8 Juxtaposition of the natural-color composite image (left) and the corresponding NDMI map (right)

2.1.8 Modification of Normalized Difference Water Index

In the study conducted by Xu (2006), it was demonstrated that the MNDWI improves the detection of surface water bodies by successfully minimizing interference caused by urban areas, plants, and soil. It was stated that, unlike NDWI, which can lead to an exaggerated detection of water areas owing to mixed signals with urban land noise, MNDWI proves to be better suited for extracting water features in areas characterized by extensive urban development.

The MNDWI is derived using the green and $SWIR_1$ bands of satellite imagery, and its calculation formula is provided in Equation 2.8. And, the juxtaposition of the natural-color composite image and the corresponding NDMI map is shown in Figure 2.9.

$$MNDWI = \frac{(Green - SWIR_1)}{(Green + SWIR_1)} \tag{2.8}$$

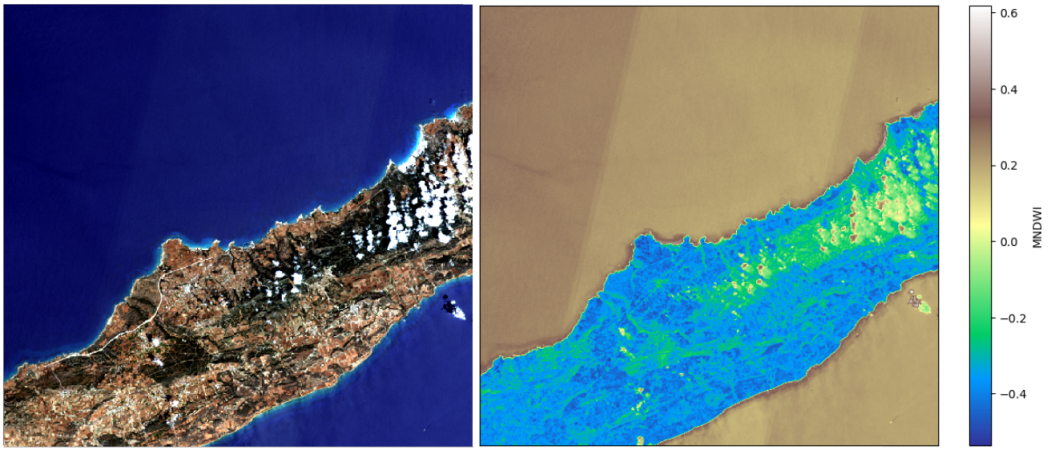


Figure 2.9 Juxtaposition of the natural-color composite image (left) and the corresponding MNDWI map (right)

2.1.9 Moisture Stress Index

The MSI is a reflectance−based index which is sensitive to variations in leaf moisture levels. Its values typically range from 0 to over 3, with green plant cover generally falling within the range of 0.4 to 2 (Sentinel Hub, 2025). The MNDWI is derived by utilizing the NIR and SWIR-1 channels of satellite imagery, and its calculation formula is provided in Equation 2.9 (Park et al., 2019).

$$MSI = \frac{(SWIR_1)}{(NIR)} \tag{2.9}$$

In the study conducted by Elhag & Bahrawi (2016), the hydrological drought indices Water Supply Vegetation Index (WSVI), SAVI, MSI, and Normalized Difference Infrared Index (NDII) were utilized to investigate the influence of salt

concentrations on crop water limitations under arid conditions.

Welikhe et al. (2017) highlighted in their study that MSI responds notably to fluctuations in soil moisture, showing higher values as soil moisture decreases. As a result of their observations, they noted that the strongest negative correlation occurred at a soil depth of 20 cm during the growing season.

The juxtaposition of the natural-color composite image and the corresponding MSI map is shown in Figure 2.10.

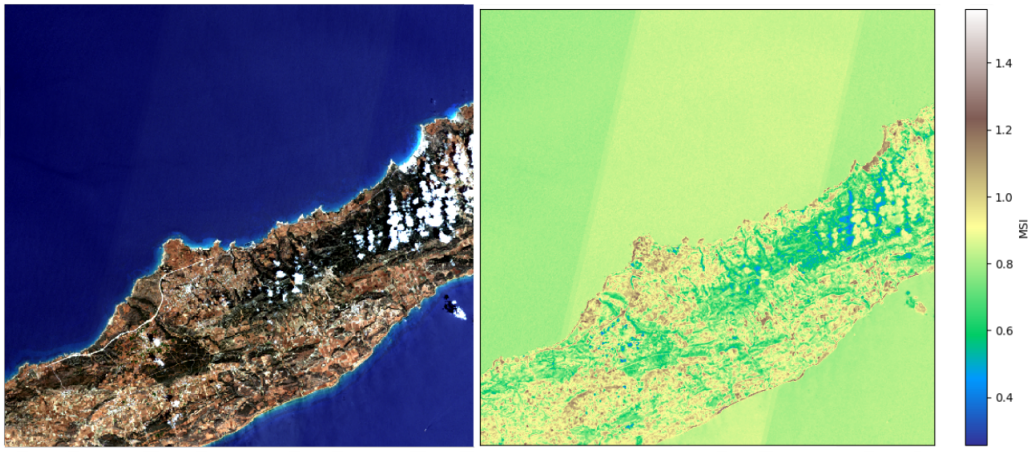


Figure 2.10 Juxtaposition of the natural-color composite image (left) and the corresponding MSI map (right)

2.2 SMI Estimation

2.2.1 Land Surface Temperature

Land surface temperature, measured in Kelvin, is a crucial geophysical parameter for global energy balance analyses and hydrologic modeling (USGS, 2025b). It is a key input for SMI estimation as it provides indirect information about soil moisture levels. According to the study conducted by Avdan & Jovanovska (2016), the first step in calculating LST involves the calculation of TOA using the TIRS-1 band of Landsat 8. The calculation formula of the transforming the pixel values of TIRS-1 to spectral radiance values is presented below as an Equation 2.10.

$$L_\lambda = M_L \cdot Q_{cal} + A_L - O_i \tag{2.10}$$

where Q_{cal} represents the pixel values of the TIRS-1 band, while M_L, A_L, and O_i are metadata parameters extracted for calibration purposes. M_L is the band-specific multiplicative rescaling factor, A_L is the band-specific additive rescaling factor, and O_i is the correction factor for the TIRS-1 band.

The second step involves converting these spectral radiance values into brightness temperatures using the formula shown in Equation 2.11.

$$BT = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)} - 273.15 \tag{2.11}$$

where K_1 and K_2 represents the band-specific thermal conversion constants.

Although brightness temperature values indicate temperature, they do not represent the actual temperature in the field. LST, which reflects the true surface temperature, is derived by applying a correction to the brightness temperature values for ground surface emissivity. The calculation of LST is presented below as an Equation 2.12.

$$LST = \frac{BT}{1 + \left[\left(\frac{\lambda \cdot BT}{\rho}\right) \ln \varepsilon_\lambda\right]} \tag{2.12}$$

where λ is the wavelength of emitted radiance (for TIRS-1, $\lambda = 10.895$), and ε_λ is the emissivity. Finally, ρ is a constant equal to $1.438 \times 10^{-2} mK$, calculated as shown in Equation 2.13.

$$\rho = \frac{h \cdot c}{\sigma} \tag{2.13}$$

where σ is the Boltzmann constant ($1.38 \times 10^{-23} J/K$), h is Planck's constant

($6.626 \times 10^{-34} Js$), and is the velocity of light ($2.998 \times 10^{8} m/s$).

This emissivity values are calculated using the NDVI threshold method.

2.2.1.1 The NDVI threshold method

To estimate LST, it is essential to determine the land surface emissivity, as it serves as a proportionality factor that adjusts black body radiance (as described by Planck's law) to predict emitted radiance. Additionally, ε_λ represents the efficiency of thermal energy transfer from the surface to the atmosphere Avdan & Jovanovska (2016). The calculation of ε_λ is presented in Equation 2.14 (Wang et al., 2015).

$$\varepsilon_\lambda = \begin{cases} \varepsilon_{s\lambda}, & NDVI < NDVI_s \\ \varepsilon_{v\lambda} P_v + \varepsilon_{s\lambda}(1 - P_v) + C_\lambda, & NDVI_s \leq NDVI \leq NDVI_v \\ \varepsilon_{v\lambda} P_v + C_\lambda, & NDVI > NDVI_v \end{cases} \tag{2.14}$$

where $\varepsilon_{v\lambda}$ represents the vegetation emissivity, while $\varepsilon_{s\lambda}$ is the soil emissivity. P_v is the proportion of vegetation, $NDVI_v$ and $NDVI_v$ are the representations NDVI values of pixels fully covered with vegetation or soil, respectively. Finally, C_λ accounts for surface roughness and is set to 0 in this study, indicating a flat surface.

The threshold values $NDVI_v$ and $NDVI_v$ are 0.5 and 0.2, respectively (Sobrino et al., 2004). The calculation formula of the proportion of vegetation is presented in Equation 2.15.

$$P_v = \left[\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s} \right]^2 \tag{2.15}$$

In this study, the surface emissivity values for different terrestrial materials Wang et al. (2015) have been used. The terrestrial materials, NDVI ranges and the surface

emissivity values are presented in Table 2.1.

Table 2.1 Emissivity values based on NDVI ranges

NDVI Range	Terrestrial Material	Emissivity
NDVI < 0	Water	0.991
$0 \leq$ NDVI < 0.2	Soil	0.996
NDVI > 0.5	Vegetation	0.973

The role of LST in SMI prediction is explained in Section 2.2.2.

2.2.2 Soil Moisture Index

In this study, SMI has been calculated using the linear correlation between LST and vegetation indices, following the methodology outlined in Zeng et al. (2004). In their repeated experiments, they observed that the NDVI-LST surface formed a typical trapezoid, as presented in Figure 2.11. Based on these observations, the proposed SMI, derived from the Ts-NDVI space, was noted to take a value of 0 along the dry edge and 1 along the wet edge.

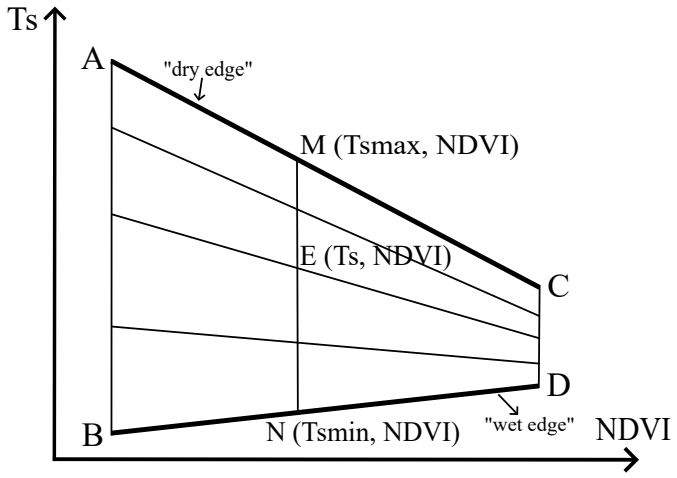


Figure 2.11 The Ts-NDVI space (Source: Zeng et al. (2004))

The formula for the SMI is presented as follows in Equation 2.16.

$$SMI = \frac{Ts_{max} - Ts}{Ts_{max} - Ts_{min}} \tag{2.16}$$

where Ts_{max}, Ts_{min} represent the maximum and minimum surface temperatures corresponding to a given NDVI value. In this study, Theil-Sen regression has been employed to obtain Ts_{max} and Ts_{min} based on the relationship between NDVI and LST. The linear models are presented in Equation 2.17 and 2.18.

$$Ts_{max} = a_1 \cdot NDVI + b_1 \tag{2.17}$$

$$Ts_{min} = a_2 \cdot NDVI + b_2 \tag{2.18}$$

In this study, SMI has been used as a key output variable to evaluate the relationship between soil moisture and various land surface factors. SMI values estimated using LST and NDVI have been utilized for both prediction and forecasting. The juxtaposition between the natural-color composite image and the corresponding NDVI, LST, and SMI maps is presented in Figure 2.12.

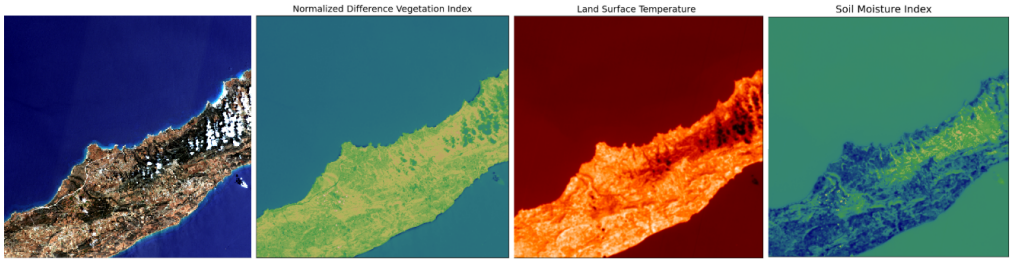


Figure 2.12 Juxtaposition of the natural-color composite image, NDVI, LST and SMI

2.3 Machine Learning Models

In this study, RF, GBM, XGBoost, and LightGBM methods have been implemented to predict the SMI, each offering distinct advantages in capturing complex relationships between land surface factors and soil moisture. For the forecasting task, the ConvLSTM model has been utilized.

2.3.1 Random Forests

RF is an ensemble learning method that improves upon traditional bagging by introducing decorrelation among the decision trees. Instead of using all predictors at each split, RF selects a random subset of predictors at each split, typically calculated as the square root of the total number of predictors. This ensures that not all trees rely on the same strongest predictor at the top split, reducing the correlation between trees. As a result, the predictions become more reliable, with lower variance compared to bagging. Unlike bagging, which averages highly correlated tree outputs, RF's decorrelation leads to better performance, particularly in datasets with many correlated predictors. Moreover, expanding the number of trees in RF does not lead to overfitting, as the error rate stabilizes when a sufficient number of trees is used (James et al., 2023).

The Random Forest algorithm operates by constructing an ensemble of decision trees (T_b) for regression or classification tasks. The steps of the algorithm for a regression task are as follows (Hastie et al., 2017).

Using the original training dataset, a bootstrap sample (Z^*) of size N is extracted. This sample is created by random sampling with replacement, ensuring diversity among the training subsets used for each tree.

For each bootstrap sample, a random forest tree is grown by recursively splitting nodes until a predefined minimum node size (n_{min}) is reached.

At each split:

i. m predictors are randomly selected from the p available predictors. The value of m is typically set to $\sqrt{p}$ for regression tasks.

ii. The best variable or split-point is chosen from the m selected predictors to minimize the impurity measure (e.g., Gini index or variance).

iii. The node is split into two child nodes, and the procedure is iteratively repeated.

After constructing B trees, the ensemble of $\{T_b\}_1^B$ is produced.

The final prediction at a new data point x is the average of predictions made by all B trees, as shown in Equation 2.19 below.

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2.19}$$

RF has been widely used in various studies for monitoring and predicting drought conditions. Some notable examples include the following: Chen et al. (2012) compared RF to ARIMA for forecasting the time series of monthly SPI, finding that RF outperformed ARIMA in both short- and long-term drought forecasting. In the study by Zhang et al. (2021), land surface temperature, vegetation indices, soil texture, geographical information, and precipitation were utilized to predict surface soil moisture (SSM). This research was grounded in on-site soil moisture observations from the International Soil Moisture Network (ISMN) and employed RF for prediction. The results demonstrated that RF effectively captured the spatio-temporal distribution of SSM on a global scale. Wang et al. (2022) analyzed the effects of multiple drought factors, including the precipitation state index (GPMI), temperature condition index (TCI), EVI, aspect, slope direction, and elevation, using RF to predict SPEI on a 6-month time scale. Zarei et al. (2023) examined the abilities of drought indices in drought assessment and determined that SPEI and Modified-SPEI indices are suitable for assessing agricultural drought conditions.

2.3.2 Gradient Boosting Machines

The gradient boosting method constructs a sequence of models that collaboratively contribute to a single predictive model. Each successive model in the sequence is trained to address the residuals from the predictions of the preceding model. This approach leverages the gradient descent algorithm to minimize a differentiable loss function, refining predictions incrementally at each step (Karaali, 2023). The steps of the algorithm are as follows (Hastie et al., 2017).

The algorithm starts by initializing the model with a constant value, $f_0(x)$, which minimizes the chosen loss function $L(y_i, \gamma)$ over all training samples. This initialization step is represented as follows in Equation 2.20.

$$f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma) \tag{2.20}$$

where y_i represents the true target value, and γ is the predicted value. The model is trained iteratively for $m = 1, 2, ..., M$ where M is the total number of boosting iterations.

For each iteration:

i. For every training sample i, calculate the pseudo-residuals r_{im} based on the gradient of the loss function L with respect to the current model predictions. The computation of the pseudo-residuals is shown in Equation 2.21.

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \tag{2.21}$$

This step captures the direction and magnitude of the error for each data point. This tree partitions the data into terminal regions R_{jm} where $j = 1, 2, ..., J_m$ (the number of terminal nodes).

ii. For each terminal region, the optimal leaf value γ_{jm} is calculated by minimizing the loss function within that region, as presented in Equation 2.22.

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right) \tag{2.22}$$

iii. The model is updated by adding the contributions of the new tree. The update rule is shown in Equation 2.23.

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right) \tag{2.23}$$

where $I\left(x \in R_{jm}\right)$ is an indicator function that equals to 1 if x belongs to region R_{jm}, and 0 otherwise.

After M iterations, the final model is the result of the last iteration, as presented Equation 2.24.

$$\hat{f}(x) = f_M(x) \tag{2.24}$$

The iterative nature of the algorithm ensures that each step incrementally improves the overall prediction accuracy.

A number of recent studies have explored the application of machine learning models in hydrological and soil moisture prediction, highlighting the effectiveness of advanced ensemble methods such as GBM and RF. The regionalization of hydrological model parameters was examined in a study by Song et al. (2022), where it was shown that the GBM model outperformed multiple linear regression (MLR) in estimating model parameters based on topographic and edaphic characteristics. For surface soil moisture mapping, RF, GBM, stacked model, and Cubist models were juxtaposed by Das et al. (2022). In the study, it was demonstrated that GBM achieved better results compared to RF and Cubist models after fine-tuning.

2.3.3 XGBoost

XGBoost, based on decision tree and gradient boosting algorithms, was initially developed by Friedman in 2002 and later presented as an advanced implementation by Chen & Guestrin (2016) (Yangın, 2019).

The objective function at the t-th iteration of XGBoost is calculated using the formula shown in Equation 2.25 (Chen & Guestrin, 2016).

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{2.25}$$

The first term of the objective function, the loss term $l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$ represents the training loss at the t-th iteration. $\hat{y}_i^{(t-1)}$ represents the prediction for the i-th data point from the model at the $(t-1)$-th iteration. $f_t(x_i)$ is the contribution of the newly added tree f_t at the t-th iteration. This term ensures the new tree corrects the residuals from the previous predictions.

The second, regularization term $\Omega(f_t)$ controls the complexity of the model. It is defined as presented in Equation 2.26.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2.26}$$

where T represents the number of leaves, while λ is a parameter that scales the penalty on the squared weights ($\|w\|^2$), which denote the values in the leaves of the tree. The parameter γ penalizes the addition of new leaves. The regularization term promotes simpler models by smoothing the final weights.

This iterative improvement and regularization are what make XGBoost robust and efficient for a wide variety of machine learning tasks. In the study by Schönauer et al. (2022), which focused on modeling spatio-temporal variations in soil moisture content, Random Forests and XGBoost achieved the highest predictive performance, with R^2 values of 0.49 and 0.51, respectively. In the comparative study conducted by Wang & Fu (2023), the performances of nine different algorithms were evaluated for soil moisture estimation. The findings showed that Random Forests achieved the highest performance, while XGBoost demonstrated the poorest performance within the scope of this study.

2.3.4 LightGBM

Developed by Ke et al. (2017), LightGBM is an ensemble learning algorithm that offers an optimized version of the gradient boosting technique. Its main objective is the significant acceleration of the training process, which often results in enhanced model

performance. Based on decision tree algorithms, LightGBM is specifically designed for efficiency and scalability when working with large-scale datasets (Rufo et al., 2021).

LightGBM employs a leaf-wise growth strategy and a histogram-based approach, which enables faster computation and lower memory usage compared to XGBoost and GBM. A key technique that enhances the performance of LightGBM is Gradient-based One-Side Sampling (GOSS). This method improves the data sampling process by leveraging the gradient values, allowing it to prioritize instances with larger gradients. GOSS selects training examples as follows (Ke et al., 2017):

i. **Ranking:** Training samples are ordered in descending order according to their absolute gradient values.

ii. **Subset Selection:** The top $\alpha \times 100\%$ of instances with the largest gradients are selected as subset A.

iii. **Random Selection:** From the remaining $(1-\alpha)\times 100\%$ of instances with smaller gradients, a random subset B is selected.

iv. **Variance Gain:** The variance gain $\tilde{V}_j(d)$ is estimated over the union of subsets using the following formula:

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-\alpha}{\beta}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-\alpha}{\beta}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right) \quad (2.27)$$

where $A_l = \{x_i \in A : x_{ij} \leq d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \leq d\}$ and $B_r = \{x_i \in B : x_{ij} > d\}$. The coefficient $\frac{1-\alpha}{b}$ is used to normalize the sum of the gradients in subset B.

This approach estimates the variance gain over a smaller instance subset, significantly reducing computational costs. By prioritizing the instances with larger gradients while maintaining randomness in selecting smaller-gradient instances, GOSS ensures computational efficiency without sacrificing model accuracy.

LightGBM has been extensively applied to soil moisture estimation, demonstrating its ability to handle complex datasets and provide accurate predictions. For instance, Li & Yan (2024) compared the performance of four tree-based ensemble models (Random Forest, XGBoost, LightGBM, and CatBoost), three deep-learning models (Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU)), and stacking models for soil moisture estimation. Among these, LightGBM achieved the best individual performance. Furthermore, Sahaar & Niemann (2024) investigated soil moisture estimation at different depths by evaluating the feed-forward artificial neural network, Random Forests, Categorical Boosting (CatBoost), LightGBM, and XGBoost. Their findings revealed that XGBoost demonstrated the highest accuracy.

2.3.5 ConvLSTM

ConvLSTM, originally developed by Shi et al. (2015), integrates the strengths of Convolutional Neural Network (CNN) and LSTM architectures. As highlighted by ElSaadani et al. (2021), CNNs excel at capturing spatial patterns, while LSTMs are specifically intended to process temporal data and learn from previous records. ConvLSTM networks combine these capabilities by incorporating recurrent layers, similar to LSTM networks, but replacing the fully connected layers in each gate with convolutional layers. This unique structure enables ConvLSTM to effectively manage spatio-temporal dependencies, making it particularly well-suited for analyzing sequential remote sensing data in the context of drought forecasting.

The detailed computations of the ConvLSTM network for time step t are presented in Equations (2.28 - 2.33) (Kartal & Şekertekin, 2022).

i. W and U represent the weight matrices for the current input x_t and the previous hidden state h_{t-1}, respectively.

ii. b is the bias term associated with each gate.

iii. The activation function σ ensures values are scaled between 0 and 1, where 0

indicates that the corresponding information will not be passed to the next state, and 1 means it will be fully retained and passed forward.

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1} + b_f) \tag{2.28}$$

$$i_t = \sigma(W_i * x_t + U_i * h_{t-1} + b_i) \tag{2.29}$$

$$\hat{C}_t = \text{relu}(W_c * x_t + U_c * h_{t-1} + b_c) \tag{2.30}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t \tag{2.31}$$

$$o_t = \sigma(W_o * x_t + U_o * h_{t-1} + b_o) \tag{2.32}$$

$$h_t = o_t \cdot \text{relu}(C_t) \tag{2.33}$$

The ConvLSTM network operates through several interconnected components. The forget gate f_t determines which information from the previous cell state C_{t-1} should be forgotten or discarded. The input gate i_t regulates the amount of new information to be added to the cell state, while the candidate cell state computes the potential updates for the cell state. The cell state C_t is then updated by combining the retained information from C_{t-1} and the new information. The output gate o_t determines the portion of the updated cell state that contributes to the hidden state, and finally, the hidden state h_t is computed based on the output gate and the updated cell state, serving as the final representation passed to the next time step.

The effectiveness of ConvLSTM in forecasting using remote sensing data has been examined in numerous studies. For example, ElSaadani et al. (2021) conducted comparison-based analysis and found that ConvLSTM outperformed CNN in predicting soil moisture. ConvLSTM models achieved an RMSE of 2.5% and mean areal correlation coefficients of 0.9 in soil moisture predictions across the study area. Similarly, Habiboullah & Louly (2023) compared two vision transformers models and a U-Net integrated ConvLSTM model for predicting soil moisture content. Notably, the ConvLSTM layers, despite having fewer parameters achieved a MAE of 0.0262, an RMSE of 0.0354, and an unbiased RMSE of 0.0343.

Bounoua et al. (2024) compared the performance of ConvLSTM and CNN-LSTM models in water stress forecasting. Their findings revealed that ConvLSTM was more effective for short sequences, whereas CNN-LSTM yielded better results for long sequences. Based on the analysis using the Crop Water Stress Index (CWSI), ConvLSTM showed better performance for short sequences (three images), with an RMSE of 0.153 compared to 0.187 for CNN-LSTM. Conversely, CNN-LSTM demonstrated superior performance for long sequences (nine images), achieving an RMSE of 0.119 compared to 0.123 for ConvLSTM.

Moreover, Liang et al. (2024) examined the impacts of climate change and human-driven actions to vegetation dynamics in the Yellow River Water Conservation Area using a modified ConvLSTM model and compared its performance with the Random Forest (RF) model. Their results demonstrated that ConvLSTM outperformed RF in capturing the effects of climate change on vegetation growth, achieving superior accuracy with R^2 of 0.99, MAE of 0.013, and RMSE of 0.018, compared to RF's corresponding values of 0.94, 0.038, and 0.045.

2.4 Evaluation Metrics

To assess the performance of the machine learning models, three evaluation metrics were utilized: R-squared (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). In this study, R^2, and the square root of the MSE (RMSE) were used to getting insights into the accuracy and reliability of the models in predicting SMI. MAE and MSE metrics were used in the training and evaluation processes of SMI forecasting model.

2.4.1 R-squared

As explained by James et al. (2023), the R^2 statistic serves as an alternative metric for assessing model fit. Representing the proportion of variance explained, R^2 always falls between 0 and 1 and remains unaffected by the scale of the target variable (y). The calculation of R^2 is presented in Equation 2.34.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.34}$$

Here, TSS represents the total sum of squares, the calculation of which is provided in Equation 2.35.

$$TSS = \sum (y_i - \bar{y})^2 \tag{2.35}$$

where $\bar{y}$ represents the sample means of y.

RSS denotes the residual sum of squares, as outlined in Equation 2.36.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.36}$$

where $\hat{y}_i$ represents the predicted value of y_i.

2.4.2 Mean Squared Error

Mean Squared Error (MSE) is a commonly used performance metric for evaluating the accuracy of model predictions. The MSE decreases as the predicted values approach the true values. The calculation formula of MSE is presented below as an Equation 2.37 (James et al., 2023).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.37}$$

2.4.3 Mean Absolute Error

MAE offers a simple explanation, as it reflects the mean size of prediction errors using the same units as the target variable. It is a robust metric for assessing model

performance when large outliers are present. Lower MAE values indicate better model accuracy. The calculation formula of MAE is presented below as an Equation 2.38 (James et al., 2023).

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.38}$$

CHAPTER THREE
DATASETS AND APPLICATION

This chapter provides a detailed description of the datasets prepared specifically for this thesis, as well as the application processes and the results obtained.

3.1 Datasets

Satellite imagery data is utilized to obtain comprehensive information from planetary surfaces and monitor various environmental processes and changes. The selection of satellite imagery or specific spectral bands relies on the context and the objectives of the study. In this study, Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) satellite imagery data have been selected due to their open-source availability and the inclusion of thermal bands, which enable the analysis of soil moisture. Furthermore, the consistent acquisition of data at regular intervals renders it highly suitable for time series analysis.

Table 3.1 Bands of the Landsat 8 satellite (Source: USGS (2019))

Band Number	Band Name	Wavelength (μm)	Spatial Resolution (m)
Band 1	Coastal/Aerosol	0.435 - 0.451	30
Band 2	Blue	0.452 - 0.512	30
Band 3	Green	0.533 - 0.590	30
Band 4	Red	0.636 - 0.673	30
Band 5	Near Infrared (NIR)	0.851 - 0.879	30
Band 6	Shortwave Infrared-1 (SWIR-1)	1.566 - 1.651	30
Band 7	Shortwave Infrared-2 (SWIR-2)	2.107 - 2.294	30
Band 8	Panchromatic	0.503 - 0.676	15
Band 9	Cirrus	1.363 - 1.384	30
Band 10	Thermal Infrared-1 (TIRS-1)	10.60-11.19	100
Band 11	Thermal Infrared-2 (TIRS-2)	11.50-12.51	100

Since its launch on February 11, 2013, Landsat 8 has been delivering valuable data for numerous applications, for example active fire detection (Schroeder et al., 2016), mapping of water use (Senay et al., 2016), mapping of land surface temperature

(Avdan & Jovanovska, 2016), and monitoring of land degradation and desertification (Kumar et al., 2022), and estimating chlorophyll-a (Cao et al., 2020), significantly contributing to various scientific and environmental studies. Landsat 8 was developed under a cooperative initiative between the National Aeronautics and Space Administration (NASA) and the U.S. Geological Survey (USGS). The temporal resolution of the satellite is 16 days and it follows Worldwide Reference System-2 (WRS-2). The specifications of Landsat 8 bands are presented in Table 3.1 (USGS, 2019). Within the scope of this study, the blue, green, red, NIR, SWIR-1, SWIR-2, TIRS-1, and TIRS-2 bands of Landsat 8 images have been utilized for the prediction process, whereas only the red, NIR, and TIRS-1 bands have been used for forecasting. The relevant bands are presented in Figure 3.1.

In this study, Landsat 8 OLI and TIRS data have been utilized for both SMI prediction in severe drought areas and SMI forecasting in agricultural areas. Additionally, SRTM 30 m DEM (ARC 1) has been employed for the prediction process.

SRTM was a collaborative radar mission conducted by NASA and the National Geospatial-Intelligence Agency (NGA) from February 11 to 22, 2000 The mission successfully mapped nearly 80% of the Earth's terrestrial surface, focusing on areas between $60°$ north and $56°$ south latitudes, with high-resolution elevation points spaced at approximately 30 meters (USGS Earth Resources Observation and Science (EROS) Center, 2025). These data have been extensively utilized in various disciplines, particularly in environmental studies and geospatial analyses, such as predicting soil moisture. The integration of SRTM data into this study has enhanced the understanding of how topographic features influence the spatial variability of soil moisture, offering meaningful understanding of the distribution and movement of moisture across different terrains. The SRTM data covering the study area used for SMI prediction is shown in Figure 3.2.

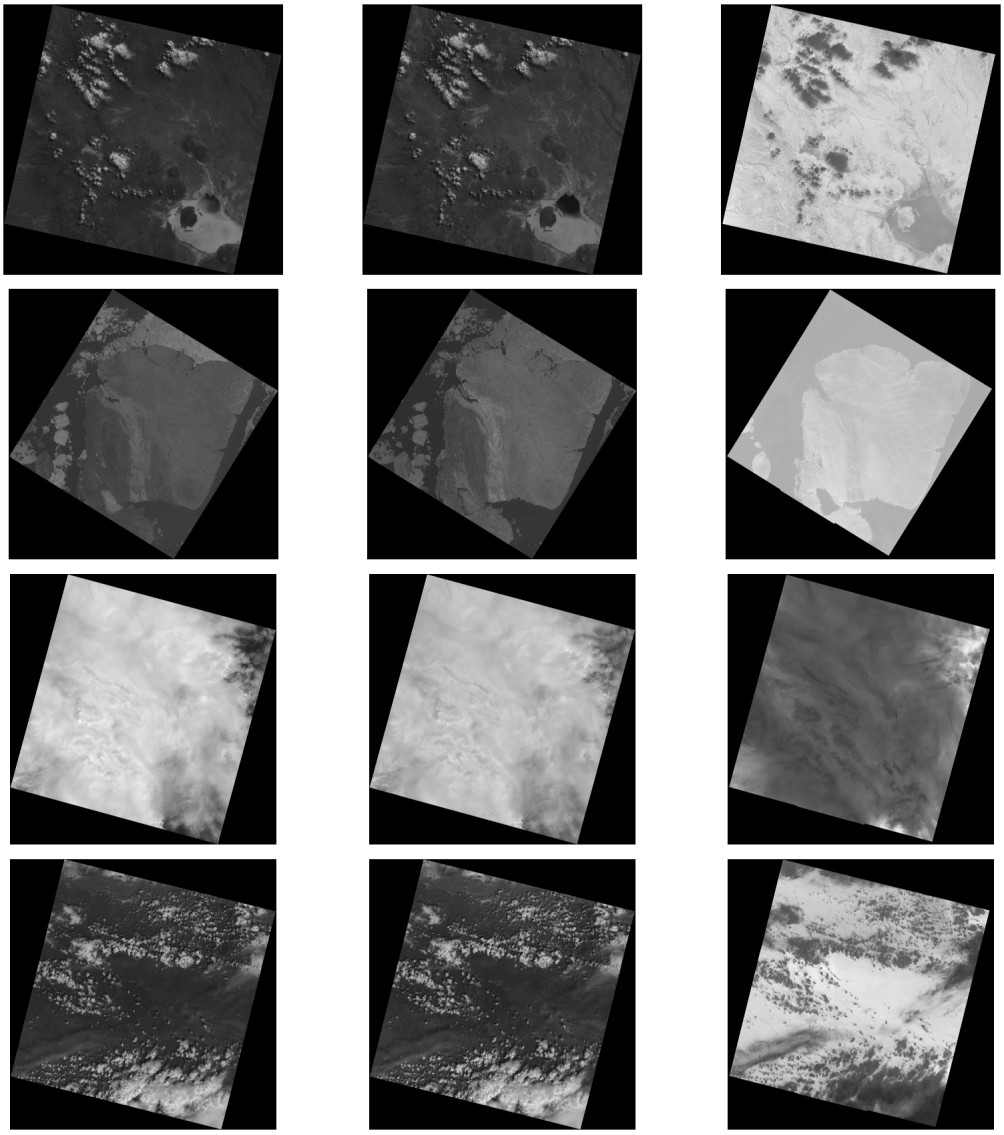


Figure 3.1 Bands of Landsat 8 images with rows corresponding to red, NIR, and TIRS-1

3.1.1 Drought Area Dataset

The development of the drought area dataset for SMI prediction has involved multiple stages, including sourcing remote sensing data to deriving relevant land surface factors. This dataset has been prepared using Landsat 8 OLI and TIRS data, as well as SRTM 30 m DEM (ARC 1) data (USGS, 2024).

The land surface factors have been grouped into four categories: vegetation, topography, water, and thermal. Table 3.2 summarizes the input variables for each category, and the data sources utilized in their obtaining. Vegetation and water indices



Figure 3.2 SRTM DEM data for the study area used in the SMI prediction process

have been derived from specific Landsat 8 bands, topographical indices have been calculated using SRTM DEM data, as detailed in Chapter 2. Thermal factors, on the other hand, have been directly obtained from the respective spectral bands of Landsat 8 imagery.

Each factor affects surface water retention in different ways. Vegetation factors restrict the loss of soil moisture, topography factors influence soil moisture content, water factors are associated with surface water dynamics, and thermal factors are connected to the evaporation of soil moisture (Park et al., 2019).

The land surface factors, input variables, and their corresponding input data are presented in Table 3.2 (Adapted from Park et al. (2019)).

All bands of 100 satellite images have been collected between 19 March 2017 and 23 June 2022, covering regions with severe drought conditions in Türkiye. The corner coordinates of study area are $34°58'34''$ N$-39°57'22''$ N and $28°00'12''$ E$-35°14'53''$ E. The coverage area of the study region is shown in Figure 3.3. After obtaining EVI, NDVI, SAVI, MSAVI, TWI, slope, aspect, NDMI, MNDWI, MSI parameters, SMI values have been estimated based on the relationship between NDVI and LST. A total of 60,641 samples have been selected for comparing the performance of RF, GBM, XGBoost and LightGBM methods for SMI prediction.



Figure 3.3 The study area of SMI prediction

Table 3.2 Land surface factors, input variables and corresponding data sources (Adapted from Park et al. (2019))

Land Surface Factors	Input Variables	Data
Vegetation	EVI	Blue, Red, NIR
	NDVI	Red, NIR
	SAVI	Red, NIR
	MSAVI	Red, NIR
Topography	TWI	SRTM DEM
	Slope	SRTM DEM
	Aspect	SRTM DEM
Water	NDMI	NIR, SWIR-1
	MNDWI	Green, SWIR-1
	MSI	SWIR-1, NIR
Thermal	NIR	NIR
	SWIR-1	SWIR-1
	SWIR-2	SWIR-2
	TIRS-1	TIRS-1
	TIRS-2	TIRS-2

The drought area dataset, including all relevant land surface factors, is publicly available and can be accessed online[2].

An agricultural area dataset has been created, as detailed below, to analyze the capability of the SMI to anticipate drought impacts in agricultural contexts, specifically in Aydın, one of Türkiye's key agricultural regions.

[2]Gök, G., Land surface factors for drought prediction, https://www.kaggle.com/datasets/gokcegok/land-surface-factors-for-drought-prediction

3.1.2 Agricultural Area Dataset

Between 27 May 2013 and 14 September 2024, the red, NIR, and TIRS-1 bands of 129 satellite images have been sequentially collected (USGS, 2024). The temporal resolution of the data is 32 days. The pilot area for this study is an agricultural region in Aydın, Türkiye, with the upper-left coordinate at 536783, 4171310 (UTM Zone 35N). Using this coordinate as a reference, 256 x 256 patches have been generated. This sequential dataset consists of 48 cloudy patches and 81 cloud-free patches. The pilot area and a sample from the dataset are shown in Figure 3.4.

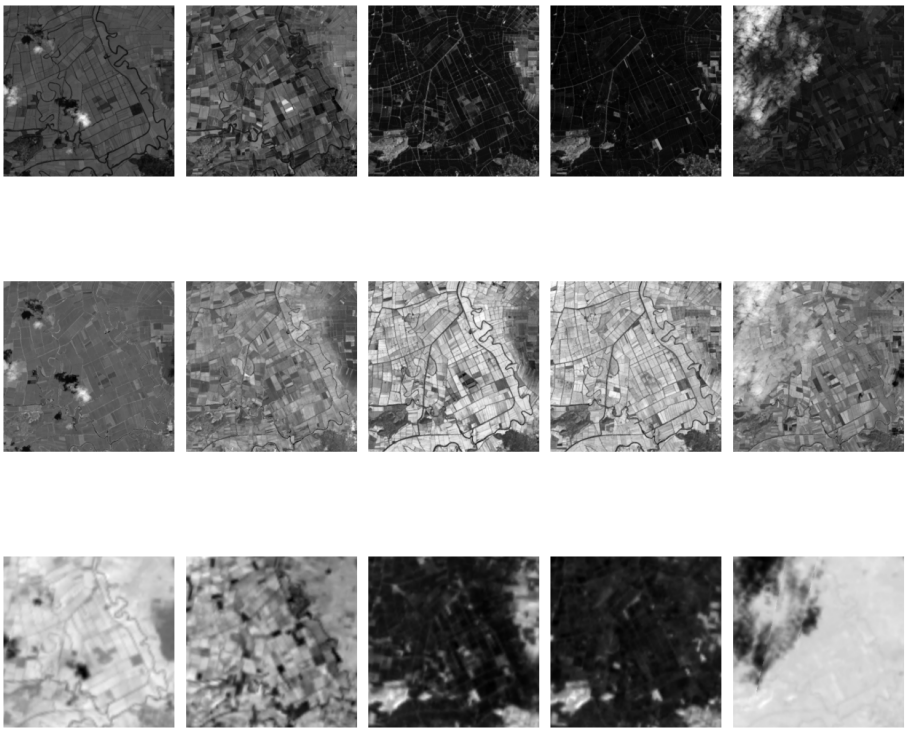


Figure 3.4 A sample of sequential data with rows corresponding to red, NIR, and TIRS-1 bands

In the forecasting process, two versions of the dataset have been utilized for comparison. A comparative analysis was conducted to evaluate two strategies: utilizing the raw data without modifications and addressing cloudy patches by substituting them with cloud-free patches retrieved from the nearest possible lag.

For SMI estimation from satellite image bands, the images have been initially used in their original dimensions (e.g., 7681 x 7801) to preserve the statistical patterns of the data.

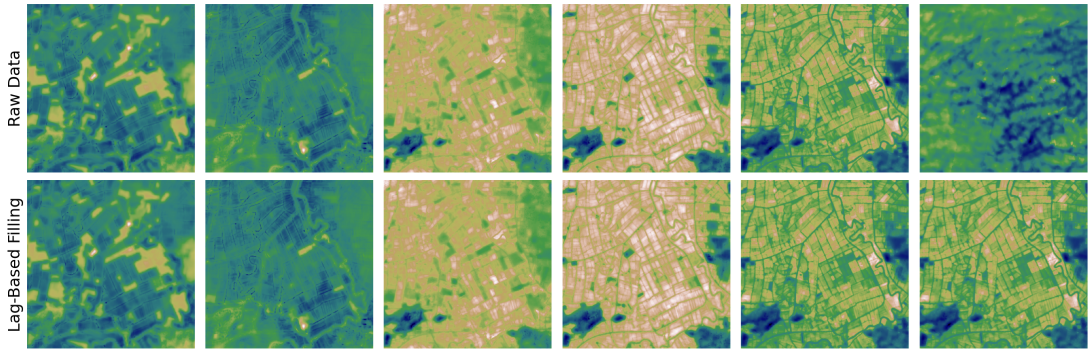


Figure 3.5 Samples from versions of agricultural area dataset

For this purpose, the pixel values of the TIRS−1 bands have been first converted to spectral radiance values using the transformation coefficients of each image. Similarly, the red and NIR bands have been converted to spectral reflectance values, and NDVI was calculated using the reflectance values of these bands. Based on NDVI values, vegetation cover ratio and land surface emissivity values have been determined. The spectral radiance values of the TIRS−1 bands have been then converted to brightness temperature values.

Finally, LST has been calculated using brightness temperature and land surface emissivity. This workflow has been fundamentally utilized in the process of deriving NDVI and LST values from satellite image bands.

The SMI has been estimated by identifying the relationship within the NDVI−LST space using Theil−Sen linear regression. Regression parameters have been determined for each image using only raw data. Once calculated, these parameters have been applied to estimate SMI values for the image patches across all two versions of the dataset. Examples from the agricultural area dataset versions, based on the estimated SMI values, are shown in Figure 3.5.

The forecasting process has been performed using the agricultural area dataset and the ConvLSTM architecture, which is particularly effective at capturing spatio-temporal dependencies.

3.2 Application

The SMI prediction focuses on severe drought areas in Türkiye, utilizing machine learning models such as RF, GBM, XGBoost, and LightGBM, with inputs derived from Landsat 8 imagery and SRTM DEM data. For SMI forecasting, a ConvLSTM architecture has been employed to predict drought conditions in agricultural areas, leveraging its ability to capture spatio-temporal dependencies in sequential remote sensing data. This two-stage approach evaluates the effectiveness of SMI in both severe drought and agricultural contexts.

The experiments for SMI prediction have been conducted using Google Colab's free-tier environment. Despite its limited computational resources, this platform adequately met the computational requirements of the study. In contrast, SMI forecasting experiments, which require more intensive computational power, have been carried out on a system equipped with a 13th Gen Intel(R) Core(TM) i7-13700H 2.40 GHz processor, 32GB of RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU, running on Windows 11.

3.2.1 SMI Prediction in Severe Drought Areas

In this study, RF, GBM, XGBoost, and LightGBM methods have been comparatively examined for predicting severe drought areas. All methods have been executed with the scikit-learn library in Python.

A sample of 60641 data points is randomly selected from the drought area dataset, which is then split into 75% training data and 25% testing data. The performance of the models has been evaluated based on R^2 and RMSE metrics for both training and test datasets.

The SMI prediction results, obtained by using the 15 input variables —NDVI, EVI, SAVI, MSAVI, slope, aspect, TWI, NDMI, MNDWI, MSI, NIR, SWIR-1, SWIR-2, TIRS-1, and TIRS-2— as detailed in Chapter 2, are presented below.

3.2.1.1 Results using Random Forests

This section showcases the outcomes achieved through the RF method for SMI prediction.

An R^2 of 0.9773 and an RMSE of 0.0279 have been achieved on the training set, while an R^2 of 0.8331 and an RMSE of 0.0757 have been achieved on the test set. The performance metrics of the model, including R^2 and RMSE for both training and test sets, are illustrated in Figure 3.6.

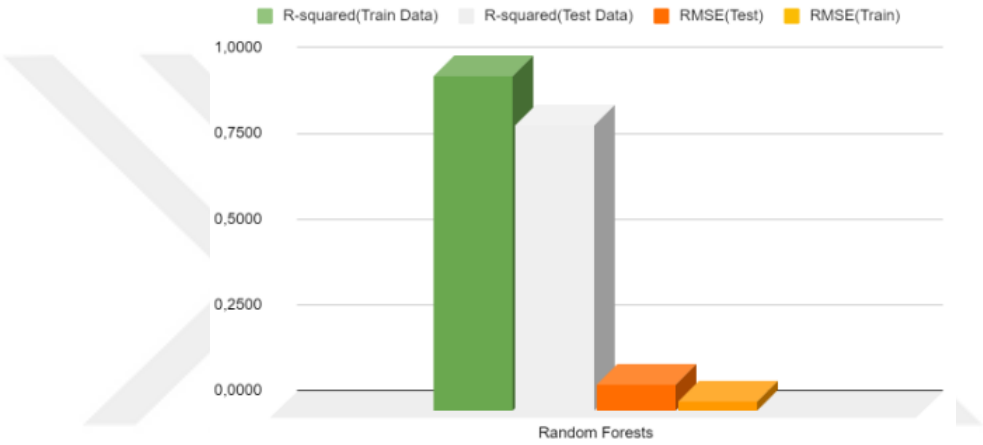


Figure 3.6 Comparison of Random Forests metrics for training and test data

The feature importance graphic is shown in Figure 3.7. When the feature importance of the 15 input variables has been analyzed, it has been observed that the three most important variables for SMI prediction using the RF method are, in order of importance, the TIRS-2 and TIRS-1 thermal bands of Landsat 8 imagery, and NDVI. These are followed, in descending order of importance, by the following variables: MNDWI, SWIR-2, MSAVI, SAVI, EVI, NIR, slope, SWIR-1, MSI, NDMI, aspect, and TWI. Notably, the topographical variables (TWI, aspect, and slope) have contributed less to the model's performance compared to the thermal, water, and vegetation indices.

To visually evaluate the performance of the RF method, the RGB versions of satellite images, the estimated SMI values considered as ground truth, the prediction results, and the error distributions are presented in Figure 3.8.

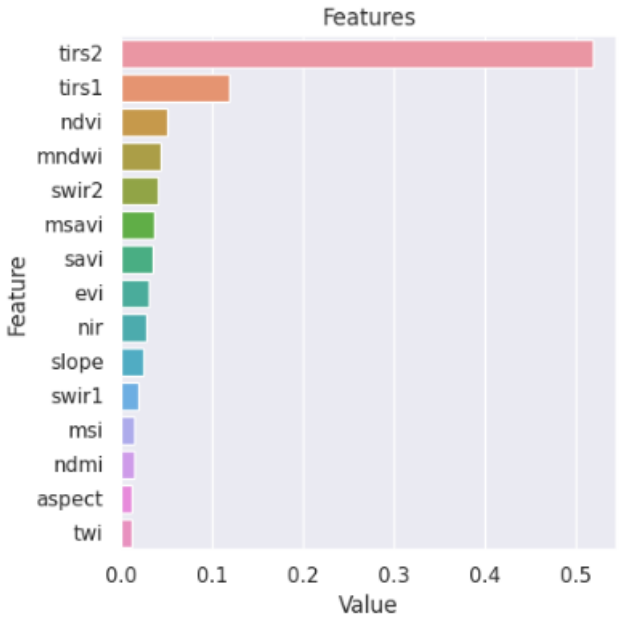


Figure 3.7 Visualization of feature importance based on Random Forests

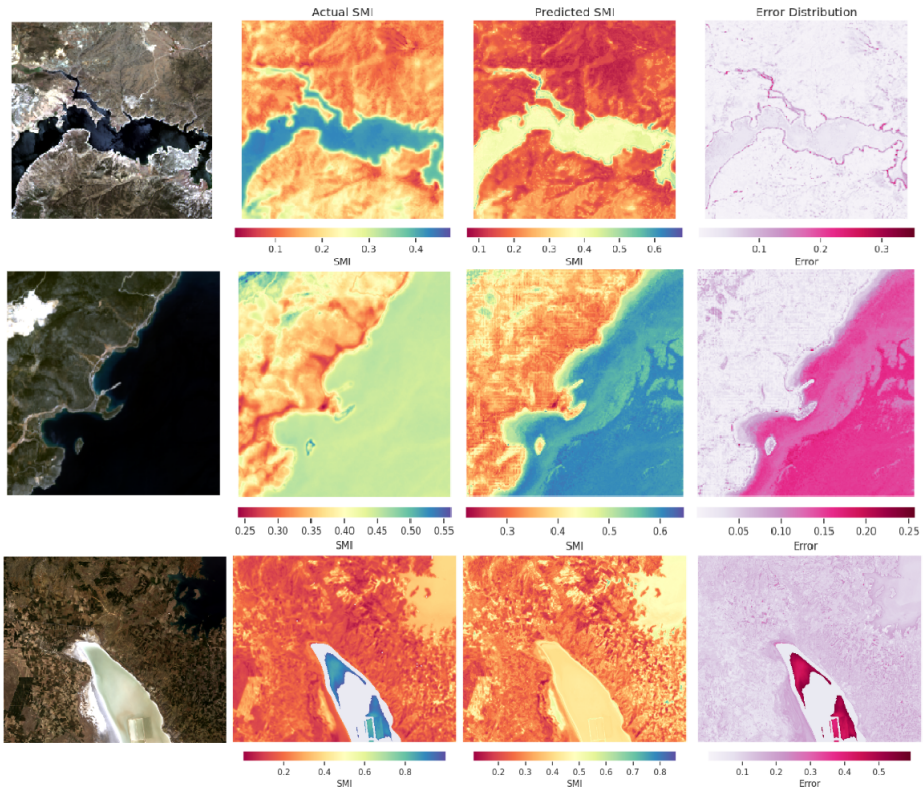


Figure 3.8 Visual representation of SMI prediction results using Random Forests

3.2.1.2 Results using GBM

GBM has been applied as an alternative to RF to evaluate the impact of boosting techniques on SMI prediction performance. This section showcases the outcomes achieved through the GBM method for SMI prediction.

GBM has demonstrated significantly lower performance compared to RF. The performance metrics of the model, including R^2 and RMSE for both training and test sets, are illustrated in Figure 3.9.

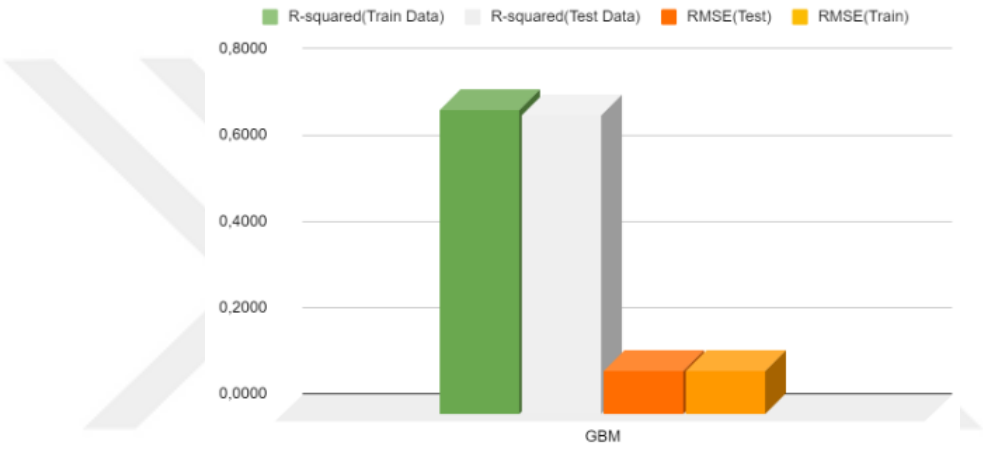


Figure 3.9 Comparison of GBM metrics for training and test data

An R^2 of 0.7049 and an RMSE of 0.1007 have been achieved on the training set, while an R^2 of 0.6948 and an RMSE of 0.1023 have been recorded on the test set. Although a decrease in performance can be observed when comparing RMSE and R^2 metrics, the closer alignment of metrics between the training and test datasets, compared to RF, suggests that the model is less prone to overfitting.

Similar to RF, the feature importance analysis of the 15 input variables has been revealed that the three most important variables for SMI prediction using the GBM method are, in order of importance, the TIRS-2 and TIRS-1 thermal bands of Landsat 8 imagery, and NDVI.

The feature importance graphic is shown in Figure 3.10. However, the ranking of the other variables in terms of feature importance has differed, as follows: SWIR-2,

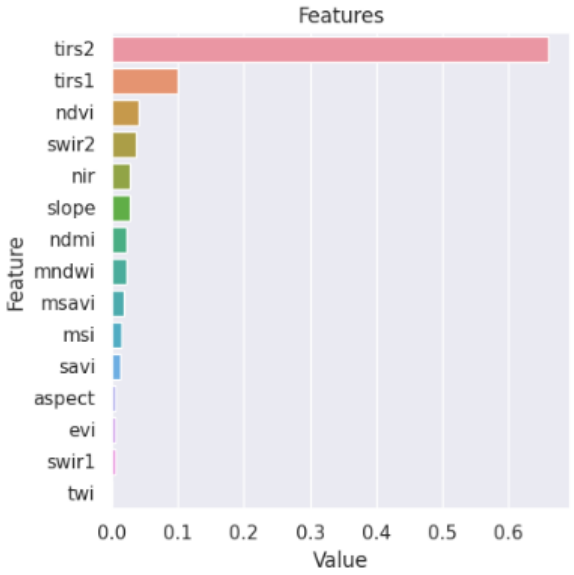


Figure 3.10 Visualization of feature importance based on GBM

NIR, slope, NDMI, MNDWI, MSAVI, MSI, SAVI, aspect, EVI, SWIR-1, and TWI. Although the slope and aspect variables have been observed to be more significant for this model, the TWI variable has been found to make no contribution.

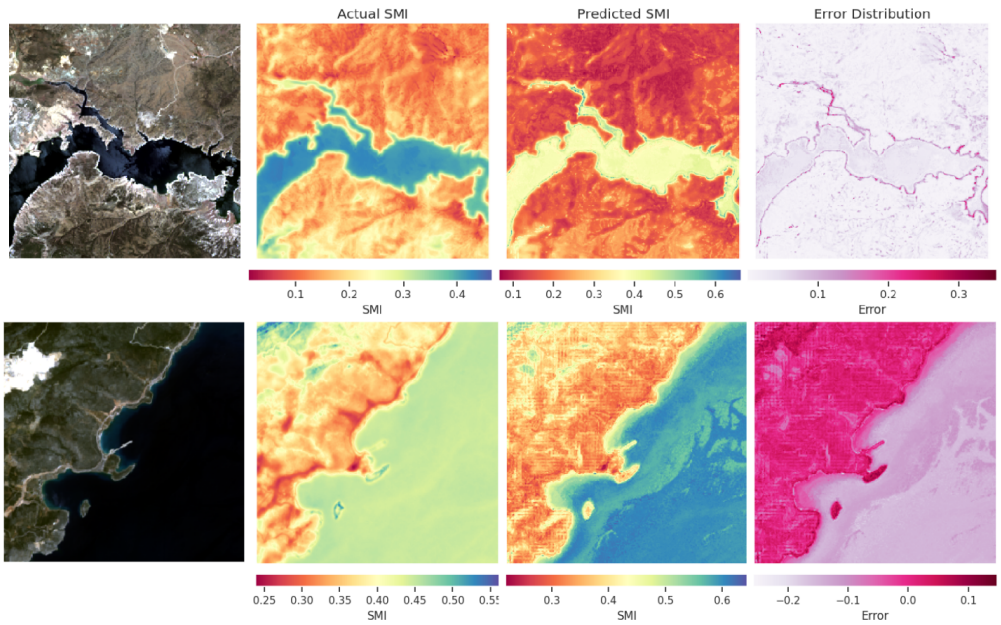


Figure 3.11 Visual representation of SMI prediction results using GBM

To visually evaluate the performance of the GBM method, the RGB versions of satellite images, the estimated SMI values considered as ground truth, the prediction results, and the error distributions are presented in Figure 3.11.

To further explore the effectiveness of boosting techniques in SMI prediction, XGBoost and LightGBM have been utilized alongside Random Forests and Gradient Boosting Machines, offering insights into their comparative performance.

3.2.1.3 Results using XGBoost

This section showcases the outcomes achieved through the XGBoost method for SMI prediction.

Although XGBoost has showed slightly lower performance than Random Forests, its results are superior to those of GBM. An R^2 of 0.8679 and an RMSE of 0.0674 have been achieved on the training set, while an R^2 of 0.7962 and an RMSE of 0.0836 have been achieved on the test set. The performance metrics of the model, including R^2 and RMSE for both training and test sets, are illustrated in Figure 3.12.

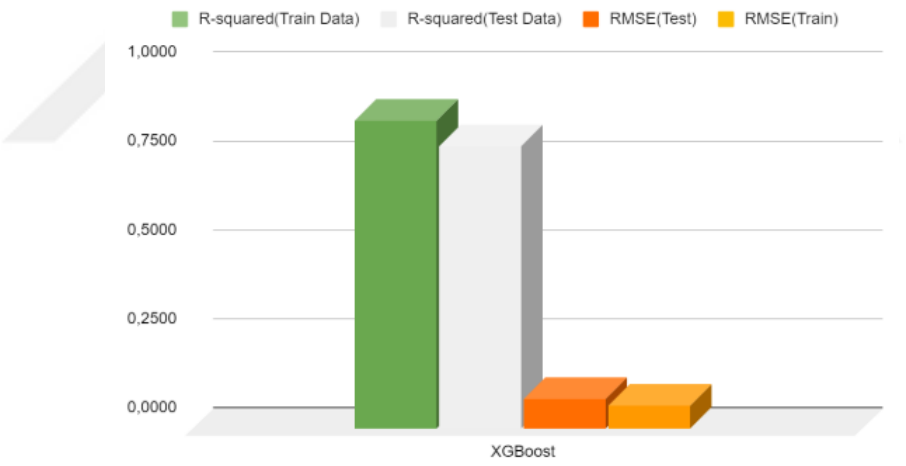


Figure 3.12 Comparison of XGBoost metrics for training and test data

The feature importance graphic is shown in Figure 3.13. Similar to RF and GBM, the feature importance analysis of the 15 input variables has revealed that the most important variable for SMI prediction using the XGBoost method was the TIRS-2 thermal band of Landsat 8 imagery. However, the feature importance rankings of the other variables differ: SWIR-2, TIRS-1, MNDWI, NDVI, slope, NIR, MSAVI, NDMI, SAVI, EVI, aspect, SWIR-1, TWI, and MSI. Notably, the MSI variable, which is of low importance in other methods, has been observed to have no impact in

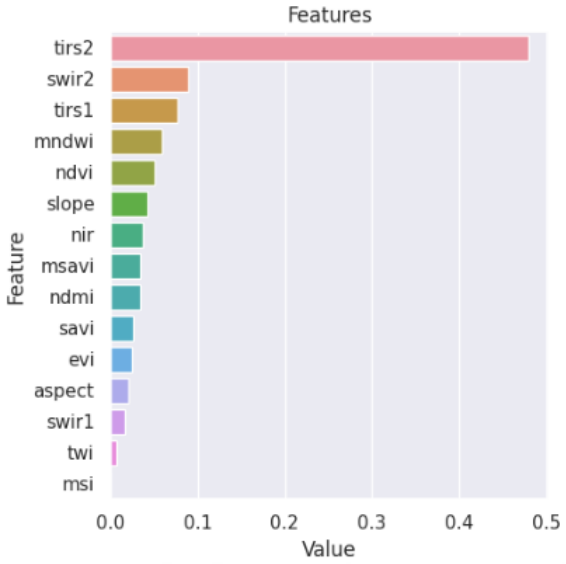


Figure 3.13 Visualization of feature importance based on XGBoost

this method.

To visually evaluate the performance of the XGBoost method, the RGB versions of satellite images, the estimated SMI values considered as ground truth, the prediction results, and the error distributions are presented in Figure 3.14.

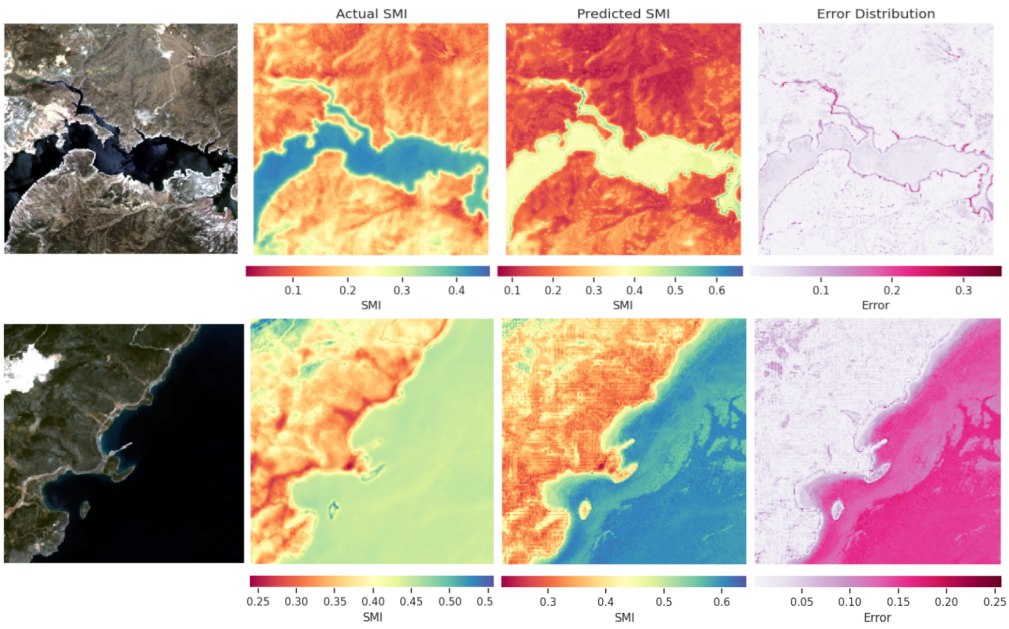


Figure 3.14 Visual representation of SMI prediction results using XGBoost

3.2.1.4 Results using LightGBM

This section showcases the outcomes achieved through the LightGBM method for SMI prediction.

Among the boosting-based models, LightGBM has demonstrated the second-best performance. The performance metrics are as follows: for the training data, an R^2 of 0.8024 and an RMSE of 0.0824 have been achieved, while for the testing data, an R^2 of 0.7748 and an RMSE of 0.0879 have been recorded. The performance metrics of the model, including R^2 and RMSE for both training and test sets, are illustrated in Figure 3.15.



Figure 3.15 Comparison of LightGBM metrics for training and test data

The feature importance graphic is shown in Figure 3.16. Differing from other models, the feature importance analysis of the 15 input variables has revealed that the most important variable for SMI prediction using the LightGBM method is the TIRS-1 thermal band of Landsat 8 imagery. Also, the feature importance rankings of the other variables differ: TIRS-2, MSAVI, SAVI, EVI, NIR, NDVI, MNDWI, slope, SWIR-1, SWIR-2, aspect, NDMI, MSI and TWI. In this method, EVI, which has been previously considered less important in other models, has been observed to be among the top five most significant variables. On the other hand, NDVI, which is consistently significant in other methods, is noted to have less importance for predictions made using this method.

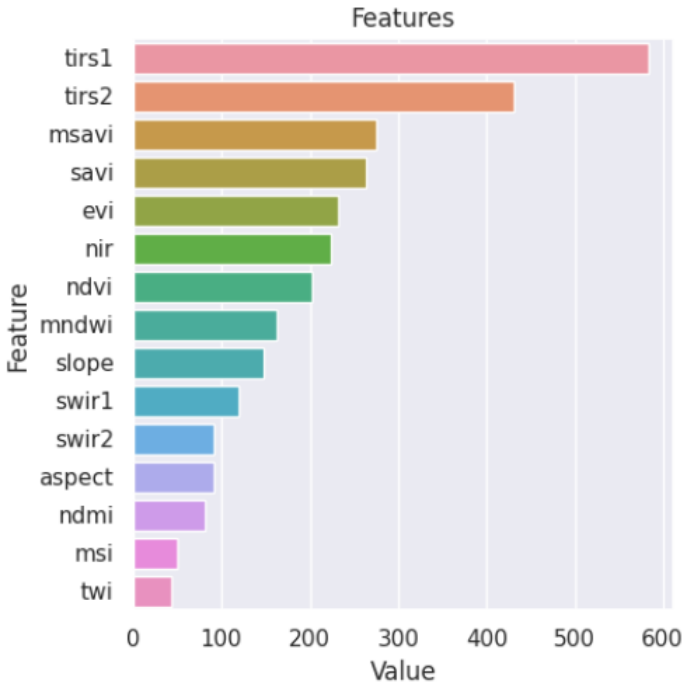


Figure 3.16 Visualization of feature importance based on LightGBM

To visually evaluate the performance of the LightGBM method, the RGB versions of satellite images, the estimated SMI values considered as ground truth, the prediction results, and the error distributions are presented in Figure 3.17.

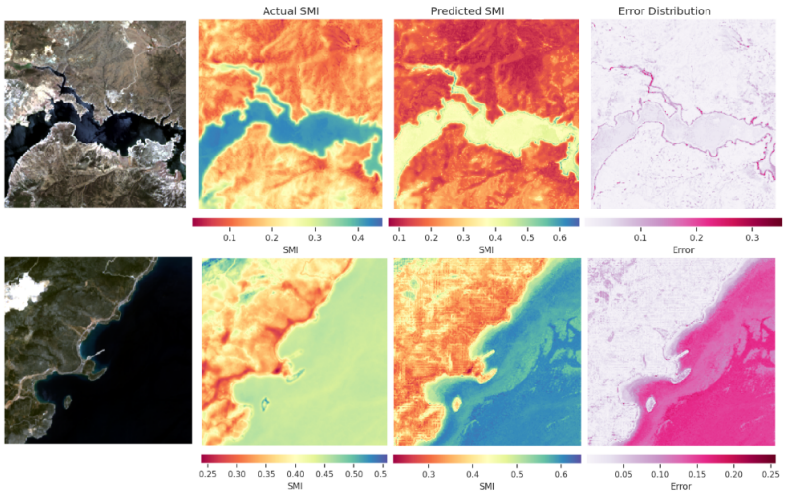


Figure 3.17 Visual representation of SMI prediction results using LightGBM

The comparison of R^2 metrics across the models is presented in Figure 3.18, while RMSE metrics are shown in Figure 3.19. And, visual comparison is presented in Figure 3.20.

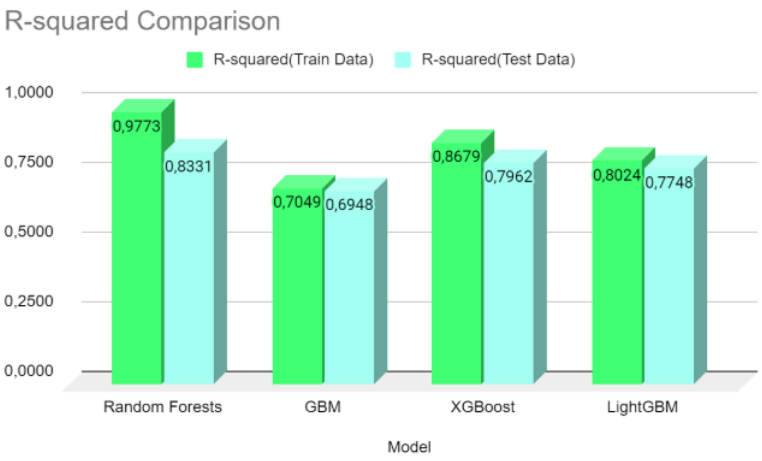


Figure 3.18 Comparison of R^2 metrics across the models

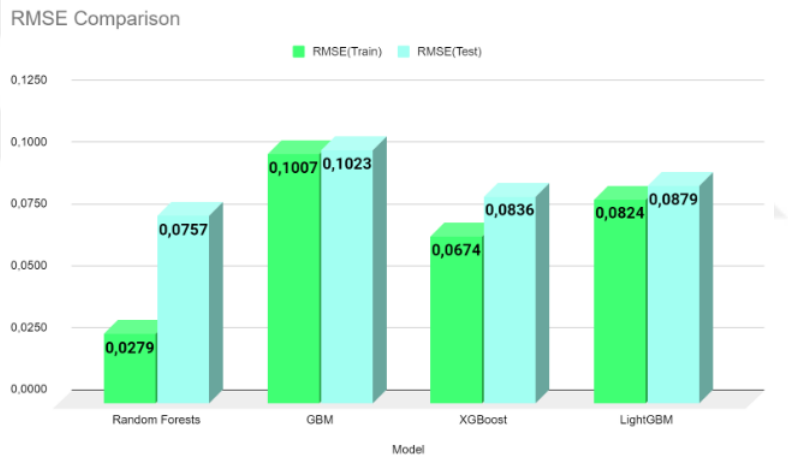


Figure 3.19 Comparison of RMSE metrics across the models

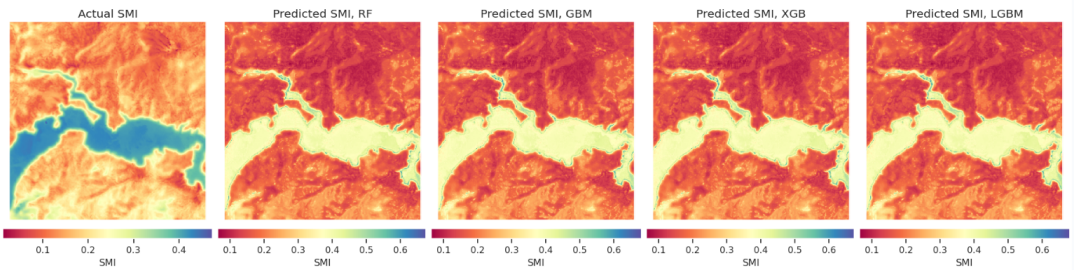


Figure 3.20 Comparison of visual outputs from SMI prediction models

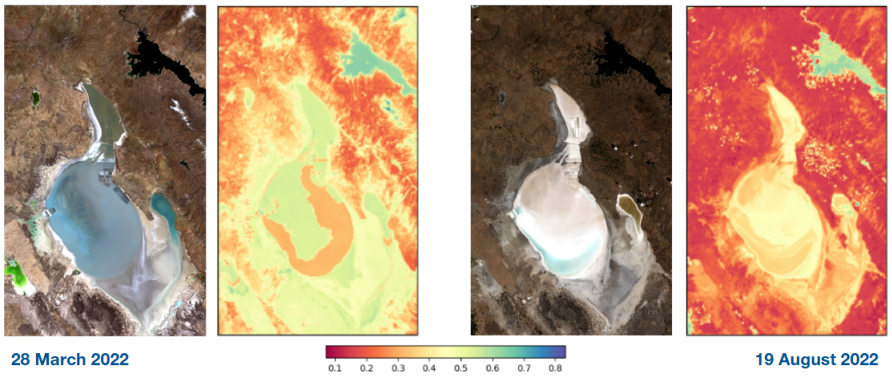


Figure 3.21 SMI prediction results using Random Forests for Lake Tuz, Türkiye

As a result, the RF algorithm has provided the best results for the dataset generated using remotely sensed data collected between 19 March 2017 and 23 June 2022, in the study area covered the southern and western parts of Türkiye. For all methods, the longwave thermal bands of Landsat 8 have been observed to play a significant role in SMI prediction. Conversely, the contribution of other land surface features has been observed to vary depending on the method used.

To further evaluate the prediction results of the RF method, an additional analysis has been conducted using satellite images of Lake Tuz from two different dates. The visual outputs of SMI prediction using RF, based on satellite images of Lake Tuz located in central Türkiye from 28 March 2022 and 19 August 2022 dates, are presented in Figure 3.21. These dates have been selected to examine whether the SMI prediction results could reflect changes in the water level of the lake. The analysis indicated that SMI values are near zero on the date associated with the lake's reduced water level, while they are significantly higher on the date of increased water levels. Therefore, while the prediction method indicates potential for further improvement, the prediction results have been considered successful.

3.2.2 SMI Forecasting in Agricultural Areas

The SMI forecasting process has been implemented using a ConvLSTM in Tensorflow 2.10.0, with Python 3.8.20. TensorFlow has been executed on the GPU using DirectML (Direct Machine Learning) API support instead of CUDA. DirectML has been chosen for its compatibility with NVIDIA GPUs, particularly on the Windows platform.

SMI forecasting has been performed using two consecutive ConvLSTM layers, each followed by batch normalization, applied to two versions of the agricultural area dataset. The network has been trained to minimize the MSE calculated on the validation dataset. While the initial plan was to train for 50 epochs, training continued as long as the validation loss decreased. Consequently, training with the raw dataset concluded at epoch 32, whereas training with the lag-filled dataset concluded at epoch 41. As presented in Table 3.3, the models have been trained with a batch size of 4, using the Adam optimizer and a learning rate of 0.001. And, the hyperparameters of ConvLSTM model are shown in Table 3.4.

Table 3.3 Hyperparameters configured for training process

Hyperparameter	**Value**
Optimizer Algorithm	Adam
Learning Rate	0.001
Batch Size	4
Epoch Number	50

Table 3.4 ConvLSTM model hyperparameters

Hyperparameter	**Value**
Number of Layers	2
Filters per Layer	32, 16
Kernel Size	7x7
Activation Function	Hard Sigmoid
Dropout Rate	0.4, 0.2
Recurrent Dropout Rate	0.2, 0.2
Padding	Same
Input Shape	(5, 256, 256, 1)

Before training, the data has been restructured to use five consecutive time steps as input for predicting the SMI value at the next time step. The training dataset spans the period from 27 May 2013 to 17 May 2021, while the test dataset covers data from 28 January 2022 to 12 July 2024.

The results obtained for the two versions of the dataset are detailed in the following sections.

3.2.2.1 Results using Raw Data

As the first step in the SMI forecasting process, the ConvLSTM network has been trained using the raw version of the sequential agricultural area dataset. As previously mentioned, the training process has been concluded after 32 epochs due to the cessation of validation loss minimization. As shown in Table 3.5, the performance metrics indicate that the model has achieved relatively similar MSE and MAE values across the training, validation, and test datasets, demonstrating a consistent predictive capability. This similarity suggests that the model generalizes well without significant overfitting. Notably, the validation and test errors are very close to each other, further confirming the model's robustness.

Table 3.5 Performance metrics for train, validation, and test data using raw data

Dataset	**MSE**	**MAE**
Training Data	0.2501	0.2140
Validation Data	0.2272	0.1985
Test Data	0.2384	0.2037

The visual outputs of SMI forecasting using raw data are shown in Figure 3.22 and Figure 3.23, which respectively present the results obtained from the training and test datasets for cases where the forecasted image corresponds to a cloudy condition.

Due to cloud cover, the SMI values that can be reliably used as ground-truth are limited to those derived from cloud-free images. Consequently, it has been concluded

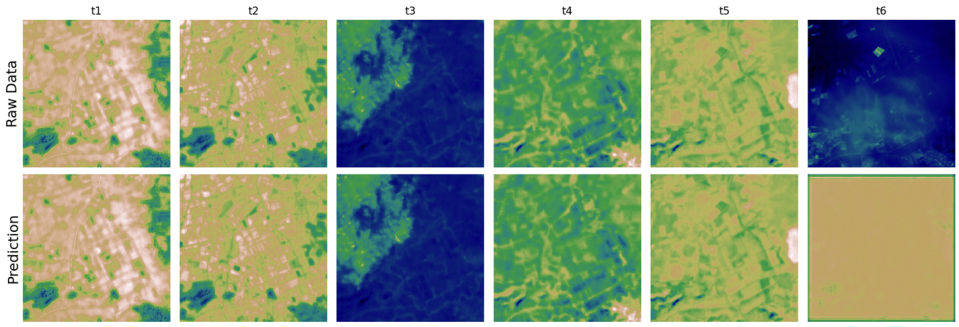


Figure 3.22 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloudy target)

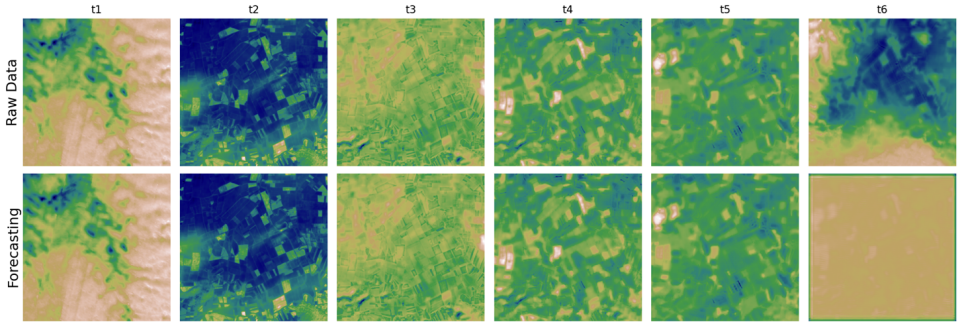


Figure 3.23 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloudy target)

that evaluating cases where the target image is cloud-free is more suitable for comparing model performance. Figure 3.24 and Figure 3.25 present the results obtained from the training and test datasets for cases where the forecasted image corresponds to a cloud-free condition.

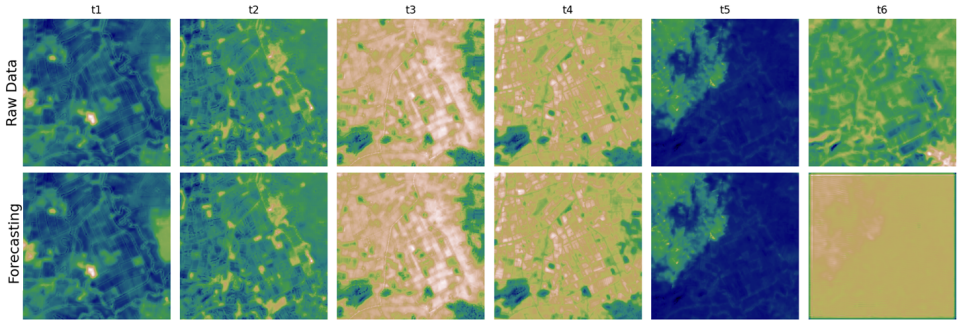


Figure 3.24 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloud-free target)

The MSE values for each time step of the test data, along with the acquisition dates of the agricultural area data and their cloudiness status, are presented in Table 3.6. As

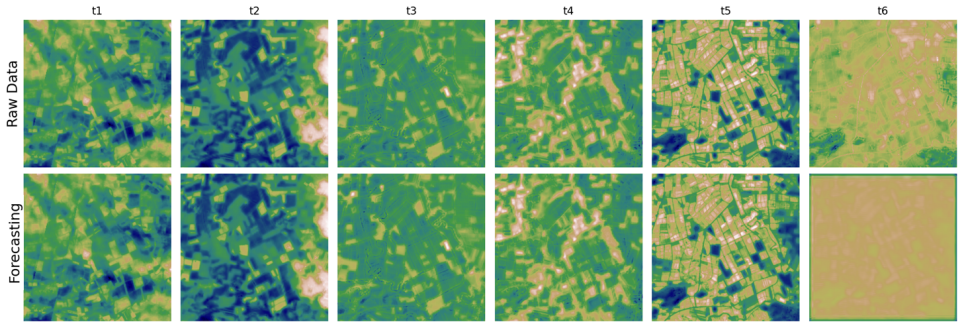


Figure 3.25 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (raw data, cloud-free target)

expected, the model has performed better on cloud-free images, reflected by a lower mean MSE value of 0.2154 compared to 0.2286 for cloudy images. This has indicated that cloud cover negatively affects the model's accuracy in predicting SMI values, underscoring the importance of using cloud-free data for reliable evaluation. The relatively small difference in MSE values between cloudy and cloud-free conditions could suggest that the model maintains a certain level of robustness under suboptimal conditions. Furthermore, visual outputs suggest that forecasting on cloud-free images may capture finer details slightly more effectively, aligning with the observed differences in MSE values and reinforcing the importance of clear conditions for accurate model performance.

Table 3.6 MSE values over time steps using raw and lag-filled data

Date	Cloudiness Status	Raw MSE	Lag-Filled MSE
07 July 2022	cloud-free	0.1724	0.1728
08 August 2022	cloud-free	0.1873	0.1870
09 September 2022	cloud-free	0.1002	0.0991
11 October 2022	cloudy	0.2198	0.0988
12 November 2022	cloud-free	0.3484	0.3543
14 December 2022	cloudy	0.1969	0.3450
15 January 2023	cloud-free	0.1797	0.1719
16 February 2023	cloud-free	0.1044	0.1068
20 March 2023	cloud-free	0.1135	0.1124
21 April 2023	cloudy	0.3407	0.1133
23 May 2023	cloudy	0.2952	0.1953
24 June 2023	cloud-free	0.1929	0.1915
26 July 2023	cloud-free	0.2696	0.2699
27 August 2023	cloud-free	0.2375	0.2380
28 September 2023	cloudy	0.2913	0.2264
30 October 2023	cloud-free	0.4205	0.4190
01 December 2023	cloudy	0.1749	0.4145
02 January 2024	cloudy	0.2828	0.0968
03 February 2024	cloud-free	0.1011	0.1059
06 March 2024	cloud-free	0.3045	0.3046
07 April 2024	cloudy	0.1725	0.3006
09 May 2024	cloudy	0.0835	0.3424
10 June 2024	cloud-free	0.3466	0.3408
12 July 2024	cloud-free	0.1526	0.1528

3.2.2.2 Results using Lag-Filled Data

The second step of the SMI forecasting process has involved training the ConvLSTM network with the lag-filled version of the sequential agricultural area dataset. As noted earlier, the training process has concluded after 41 epochs due to the halt in validation loss minimization. As illustrated in Table 3.7, similar to the results obtained using raw data, the performance metrics indicate that the model trained with lag-filled data has achieved relatively similar MSE and MAE values across the training, validation, and test datasets. These results confirm that the model generalizes well and is robust.

Table 3.7 Performance metrics for train, validation, and test data using lag-filled data

Dataset	MSE	MAE
Training Data	0.2414	0.2084
Validation Data	0.2318	0.2133
Test Data	0.2465	0.2105

When these error values have been compared with those obtained using raw data, it was noted that the errors are lower for the training data but higher for the validation and test datasets. This could potentially be explained by the manipulation of the statistical diversity of the raw data caused by lag filling, which may have increased the model's tendency to overfitting.

The visual outputs of SMI forecasting using lag-filled data are shown in Figure 3.26, Figure 3.27, which respectively present the results obtained from the training and test datasets for cases where the forecasted image corresponds to a cloudy condition. And, the visual outputs of SMI forecasting using lag-filled data for cases where the forecasted image corresponds to a cloud-free condition are shown in Figure 3.28, Figure 3.29, which respectively present the results obtained from the training and test datasets.

The MSE values for each time step of the test dataset, together with the acquisition dates of the prediction data and their respective cloudiness status, are shown in

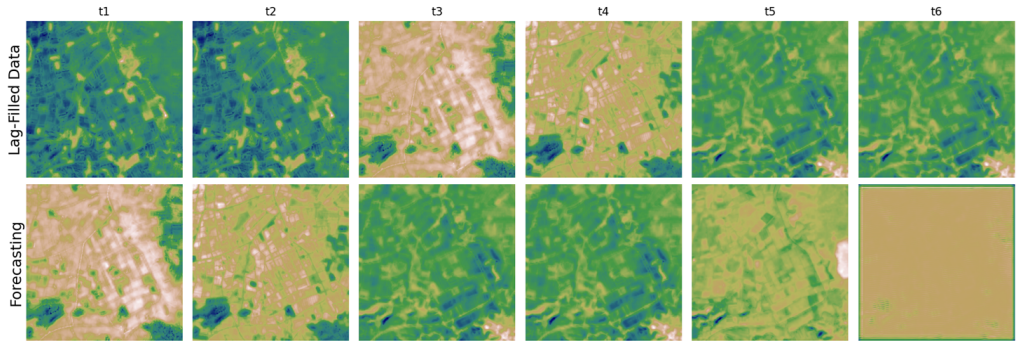


Figure 3.26 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloudy target)

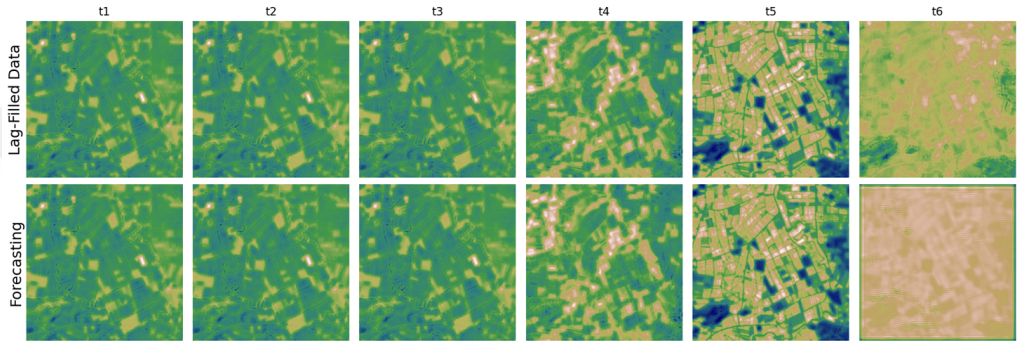


Figure 3.27 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloudy target)

Table 3.6. When compared to the results of raw data forecasting, the lag-filled dataset has shown a slight improvement in model performance for cloud-free images, with a mean MSE value of 0.2151 compared to 0.2154. However, for cloudy images, the lag-filled dataset results in a higher mean MSE of 0.2370 compared to 0.2286 observed with raw data forecasting. This suggests that while lag-filling can enhance performance for cloud-free conditions by addressing data gaps, it may introduce artifacts or reduce the statistical diversity of the dataset, possibly impacting the model's capability to generalize effectively under cloudy conditions.

A comparison of the MSE values for each time step in the test data between the raw data and the lag-filled data does not clearly indicate the superiority of one over the other. Although better performance has been observed using the lag-filled data in 56% of the actually cloudy images, the restricted availability of these images (only 9) reduces the reliability of this conclusion.

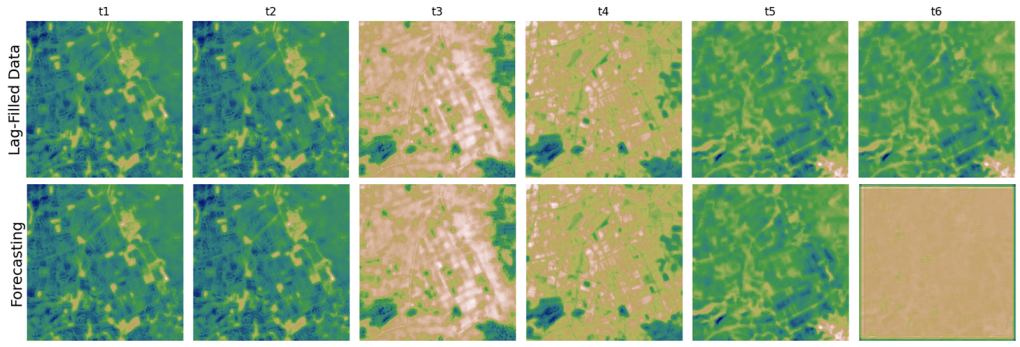


Figure 3.28 Comparison of training data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloud-free target)

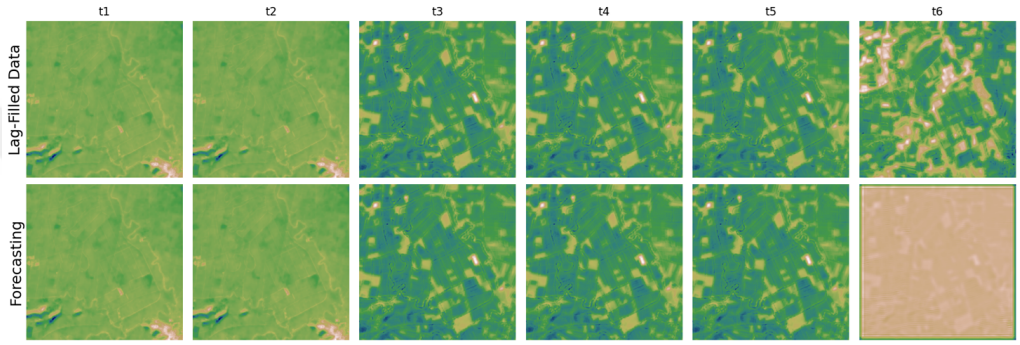


Figure 3.29 Comparison of test data and forecasting results for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset (lag-filled, cloud-free target)

Although addressing data loss caused by cloudy images through lag-filling with the nearest lag does not appear to provide a significant improvement in this context, a visual comparison (Figure 3.30) suggests that the lag-filled version captures details slightly better. Nevertheless, it is well-known that cloud-free conditions are critical for calculating indices obtained through remotely sensed data, as demonstrated in the analyses conducted in this study. Therefore, for future research, exploring alternative approaches, such as cloud removal techniques, to address cloud cover in images is worth considering.

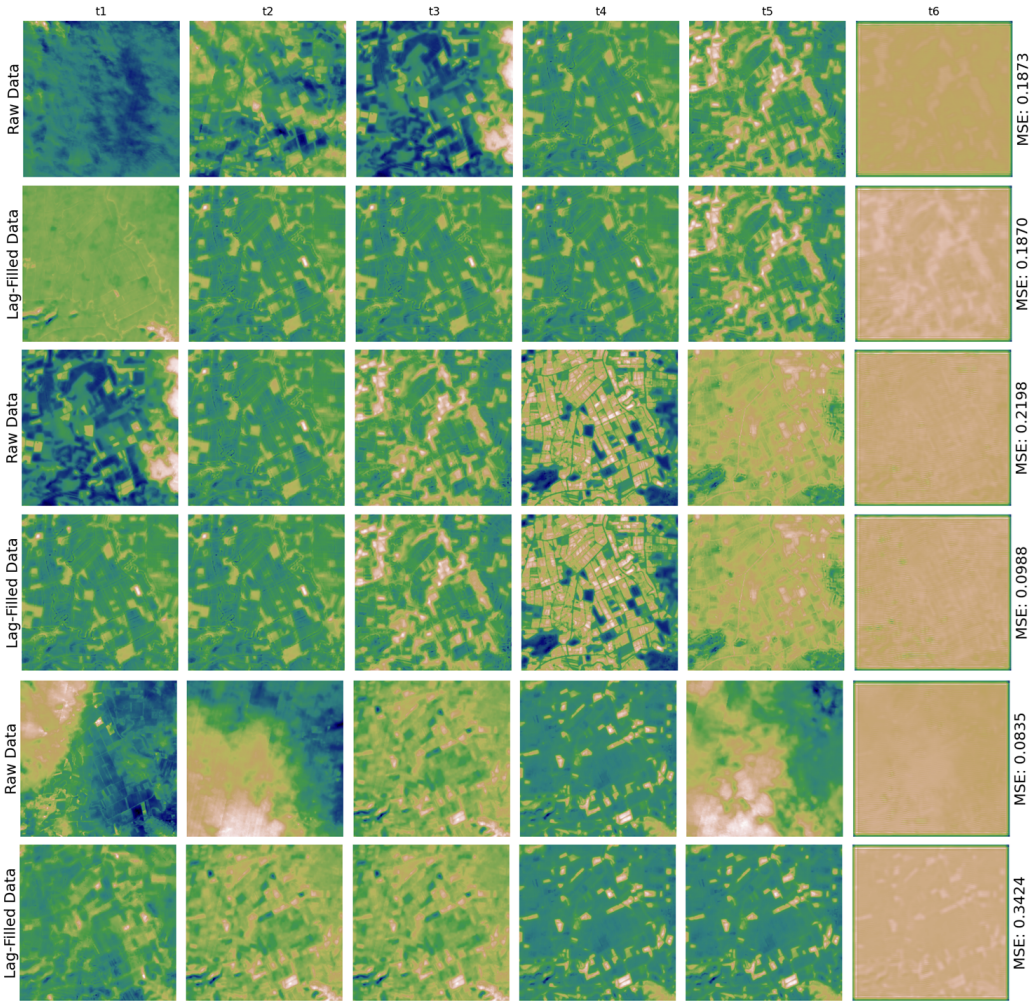


Figure 3.30 Comparison of forecasting results using raw data and lag-filled data for consecutive time steps ($t1$ to $t6$), where each t represents a sequential time step in the dataset

CHAPTER FOUR
CONCLUSION

Drought is a complex and chaotic natural disaster, characterized by its slow onset and cumulative long-term adverse environmental, social, and economic impacts on a global scale. The inherent complexity of drought poses challenges for both its precise definition and accurate forecasting (Çamalan et al., 2017). Wilhite & Glantz (1985) argue that the lack of a universally recognized definition for drought arises from its varied impacts depending on the context. Through their analysis of more than 150 definitions, the study categorized drought into four distinct types: meteorological, agricultural, hydrologic, and socio-economic. Given this complexity, robust and scalable methods are essential for improving our ability to predict and forecast drought impacts, particularly in vulnerable regions like Türkiye.

Türkiye's geographical location within the dry-summer subtropical macro-climate zone contributes to various climate-related challenges, both under current climate conditions and in light of potential future changes. These factors make the country particularly vulnerable to medium to high climate risks (Türkeş, 2020). Based on the Türkiye Desertification Risk Map from Karagöz et al. (2017), 12.7% of the country's land is identified as having low desertification risk, 53.2% as moderate risk, and 25.5% as high risk. The escalating desertification and drought risks in Türkiye underline the need for innovative approaches to monitor and predict drought conditions effectively.

Thanks to the comprehensive information, remotely-sensed data has been extensively applied in numerous researches related to drought monitoring, prediction, and forecasting. It has been extensively utilized in drought monitoring, prediction, and forecasting. Complementing this, machine learning technologies have recently gained prominence in modeling and understanding drought characteristics. These algorithms are particularly effective for drought prediction as they can integrate large volumes of data derived from multiple sources, including satellite images, meteorological stations, and soil monitoring devices. By identifying patterns and

relationships that may not be immediately apparent to humans, machine learning models can provide valuable insights. Moreover, they utilize past data to generate accurate forecasts about future drought conditions, enabling better preparedness and mitigation of drought impacts on agriculture, water supplies, and ecosystems.

This thesis aims to design and assess an SMI model for predicting severe drought conditions in Türkiye and to assess its applicability in forecasting agricultural droughts. The first part of the study focuses on modeling and predicting SMI in regions with significant drought risk using machine learning models, including RF, GBM, XGBoost, and LightGBM. These predictions have been conducted for areas known to experience severe drought using indices such as the NDVI and LST derived from Landsat 8 satellite images and SRTM DEM (30 m) data.

As a consequence of the model comparisons, the RF algorithm has delivered the best performance for the drought area dataset, which was generated using remotely sensed data collected between 19 March 2017 and 23 June 2022, in the study area covered the southern and western parts of Türkiye. An R^2 of 0.9773 and an RMSE of 0.0279 have been achieved on the training set, while an R^2 of 0.8331 and an RMSE of 0.0757 have been achieved on the test set. For all methods, the longwave thermal bands of Landsat 8 have been observed to have a major impact in SMI prediction. Conversely, the contribution of other land surface features has been observed to vary depending on the method used.

To further evaluate the prediction results of the RF method, an additional analysis has been conducted using satellite images of Lake Tuz from two different dates. The analysis indicated that SMI values are near zero on the date associated with the lake's reduced water level, while they are significantly higher on the date of increased water levels. Therefore, while the prediction method indicates potential for further improvement, the prediction results have been considered successful.

Building upon these results, the second part of the study explores the use of the SMI forecasting in agricultural areas. A forecasting model has been developed using

the ConvLSTM architecture, leveraging its capability to model spatio-temporal patterns effectively. This model has been implemented to an agricultural area dataset from Aydın, one of Türkiye's key agricultural regions. The ConvLSTM model has been trained using sequential data to forecast SMI values, with a multi-step prediction approach employing five time steps as input to forecast the sixth step.

For investigating the impact of cloud cover the forecasting process has been employed for two versions of agricultural area dataset: raw data and lag-filled data. When compared to the results of raw data forecasting, the lag-filled dataset has shown a slight improvement in model performance for cloud-free images, with a mean MSE value of 0.2151 compared to 0.2154. However, for cloudy images, the lag-filled dataset results in a higher mean MSE of 0.2370 compared to 0.2286 observed with raw data forecasting. This suggests that while lag-filling can enhance performance for cloud-free conditions by addressing data gaps, it may introduce artifacts or reduce the statistical diversity of the dataset, potentially impacting the model's capability to generalize effectively under cloudy conditions.

Although addressing data loss caused by cloudy images through lag-filling with the nearest lag does not appear to provide a significant improvement in this context, a visual comparison suggests that the lag-filled version captures details slightly better. Nevertheless, it is well-known that cloud-free conditions are critical for calculating indices obtained through remotely sensed data, as demonstrated in the analyses conducted in this study. Therefore, for future research, exploring alternative approaches to handle cloud cover, such as cloud removal techniques, as well as investigating other forecasting methods beyond ConvLSTM or implementing fine-tuning on the existing model, could potentially lead to further improvements.

This two-stage approach has provided a comprehensive framework for analyzing the effectiveness of the SMI model in both severe drought areas and agricultural contexts. By leveraging machine learning techniques for prediction and deep learning architectures for forecasting, the study has demonstrated the adaptability and utility of the SMI across varying environmental conditions. While the results underscore the

importance of cloud-free data for reliable performance, the analyses also highlight areas for potential improvement, such as exploring advanced cloud removal techniques and fine-tuning existing models to enhance their robustness. This dual approach not only contributes to the examining of drought dynamics but also presents actionable strategies for improving drought monitoring and mitigation efforts in diverse regions.

REFERENCES

AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., & Hain, C. R. (2015). Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics*, *53*(2), 452–480. https://doi.org/10.1002/2014RG000456.

American Meteorological Society (AMS) (2025). *Drought*. https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/archive-statements-of-the-ams/drought/. Accessed 2025, January 25.

Avdan, U., & Jovanovska, G. (2016). Algorithm for automated mapping of land surface temperature using Landsat 8 satellite data. *Journal of Sensors*, *2016*(1), 1480307. https://doi.org/10.1155/2016/1480307.

Bhattacharya, S., Halder, S., Nag, S., Roy, P. K., & Roy, M. B. (2021). *Assessment of drought using multi-parameter indices*. In Roy, P. K. and Roy, M. B. and Pal, S. (Eds.), *Advances in Water Resources Management for Sustainable Use* (pp. 243–255). Springer, Singapore. https://doi.org/10.1007/978-981-33-6412-7_18.

Bojer, A. K., Biru, B. H., Al-Quraishi, A. M. F., Debelee, T. G., Negera, W. G., Woldesillasie, F. F., & Esubalew, S. Z. (2024). Machine learning and remote sensing based time series analysis for drought risk prediction in Borena Zone, Southwest Ethiopia. *Journal of Arid Environments*, *222*, 105160. https://doi.org/10.1016/j.jaridenv.2024.105160.

Bounoua, I., Saidi, Y., Yaagoubi, R., & Bouziani, M. (2024). Deep learning approaches for water stress forecasting in arboriculture using time series of remote sensing images: Comparative study between ConvLSTM and CNN-LSTM models. *Technologies*, *12*(6), 77. https://doi.org/10.3390/technologies12060077.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sensing of Environment*, *248*, 111974. https://doi.org/10.1016/j.rse.2020.111974.

Çamalan, G., Akgündüz, S., Ayvacı, H., Çetin, S., Arabacı, H., & Coşkun, M. (2017). *SPEI indisine göre Türkiye geneli kuraklık değişim ve eğilim projeksiyonları* [Conference proceedings]. In *IV. Türkiye İklim Değişikliği Kongresi, (TİKDEK)*, İstanbul, Türkiye. https://www.mgm.gov.tr/FILES/genel/makale/kuraklikprojeksiyon.pdf.

Chen, J., Li, M., & Wang, W. (2012). Statistical uncertainty estimation using random forests and its application to drought forecast. *Mathematical Problems in Engineering*, *2012*(1), 915053. https://doi.org/10.1155/2012/915053.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* [Conference proceedings]. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794)., San Francisco, CA, USA. ACM. https://doi.org/10.1145/2939672.2939785.

Das, A. C., Shahriar, S. A., Chowdhury, M. A., Hossain, M. L., Mahmud, S., Tusar, M. K., Ahmed, R., & Salam, M. A. (2023). Assessment of remote sensing-based indices for drought monitoring in the north-western region of Bangladesh. *Heliyon*, *9*(2), e13016. https://doi.org/10.1016/j.heliyon.2023.e13016.

Das, B., Rathore, P., Roy, D., Chakraborty, D., Jatav, R. S., Sethi, D., & Kumar, P. (2022). Comparison of bagging, boosting and stacking algorithms for surface soil moisture mapping using optical-thermal-microwave remote sensing synergies. *CATENA*, *217*, 106485. https://doi.org/10.1016/j.catena.2022.106485.

Elhag, M., & Bahrawi, J. (2016). Soil salinity mapping and hydrological drought indices assessment in arid environments based on remote sensing techniques. *Geoscientific Instrumentation, Methods and Data Systems Discussions*, 1-31. https://doi.org/10.5194/gi-2016-39.

ElSaadani, M., Habib, E., Abdelhameed, A. M., & Bayoumi, M. (2021). Assessment of a spatiotemporal deep learning approach for soil moisture prediction and filling the gaps in between soil moisture observations. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.636234.

Gu, Z., Zeng, Z., Shi, X., Yu, D., Zheng, W., Zhang, Z., & Hu, Z. (2009). Estimating models of vegetation fractional coverage based on remote sensing images at different radiometric correction levels. *Frontiers of Forestry in China*, *4*(4), 402–408. https://doi.org/10.1007/s11461-009-0057-8.

Habiboullah, A., & Louly, M. (2023). Soil moisture prediction using NDVI and NSMI satellite data: ViT-based models and ConvLSTM-based model. *SN Computer Science*, *4*, 140. https://doi.org/10.1007/s42979-022-01554-7.

Han, P., Wang, P. X., Zhang, S. Y., & Zhu, D. H. (2010). Drought forecasting based on the remote sensing data using ARIMA models. *Mathematical and Computer Modelling*, *51*(11), 1398-1403. https://doi.org/10.1016/j.mcm.2009.10.031.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7.

Huang, F., Zhang, Y., Zhang, Y., Shangguan, W., Li, Q., Li, L., & Jiang, S. (2023). Interpreting Conv-LSTM for spatio-temporal soil moisture prediction in China. *Agriculture*, *13*(5), 971. https://doi.org/10.3390/agriculture13050971.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python* (1st ed.). Springer. https://doi.org/10.1007/978-3-031-38747-0.

Karaali, B. (2023). *Makine öğrenmesi algoritmaları ile rüzgar türbinlerinin güç tahmini*. [Master's Thesis]. Dokuz Eylül University.

Karagöz, A., Tercan, E., Erpul, G., Türkeş, M., Dengiz, O., Doğan, O., & Öztaş, T. (2017). *Türkiye çölleşme modeli teknik özet*. https://webdosya.csb.gov.tr/db/cem/icerikler/teknik-ozet-tr-20211105123742.pdf.

Kartal, S., & Şekertekin, A. (2022). Prediction of MODIS land surface temperature using new hybrid models based on spatial interpolation techniques and deep learning models. *Environmental Science and Pollution Research*, *29*, 67115–67134. https://doi.org/10.1007/s11356-022-20572-9.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree* [Conference proceedings]. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, volume 30, (pp. 3149 – 3157)., Long Beach, California, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kopecký, M., Macek, M., & Wild, J. (2021). Topographic wetness index calculation guidelines based on measured soil moisture and plant species composition. *Science of The Total Environment*, *757*, 143785. https://doi.org/10.1016/j.scitotenv.2020.143785.

Kumar, B., Babu, K., Anusha, B., & Rajasekhar, M. (2022). Geo-environmental monitoring and assessment of land degradation and desertification in the semi-arid regions using Landsat 8 OLI / TIRS, LST, and NDVI approach. *Environmental Challenges*, *8*, 100578. https://doi.org/10.1016/j.envc.2022.100578.

Li, M., & Yan, Y. (2024). Comparative analysis of machine-learning models for soil moisture estimation using high-resolution remote-sensing data. *Land*, *13*(8), 1331. https://doi.org/10.3390/land13081331.

Liang, Z., Sun, R., & Duan, Q. (2024). Attribution of vegetation dynamics in the Yellow River water conservation area based on the deep ConvLSTM model. *Remote Sensing*, *16*(20), 3875. https://doi.org/10.3390/rs16203875.

Liu, Y., Qian, J., & Yue, H. (2021). Comparison and evaluation of different dryness indices based on vegetation indices-land surface temperature/albedo feature space. *Advances in Space Research*, *68*(7), 2791-2803. https://doi.org/10.1016/j.asr.2021.05.007.

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles,

D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.

O'Callaghan, J. F., & Mark, D. M. (1984). The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics, and Image Processing*, *28*(3), 323-344. https://doi.org/10.1016/S0734-189X(84)80011-0.

Öztürk, D., & Kılıç, F. (2013). *Drenaj ağının çıkarımı ve topografik nemlilik indeksinin belirlenmesinde D8 ve D-inf algoritmalarının karşılaştırmalı analizi* [Conference proceedings]. In *TMMOB Coğrafi Bilgi Sistemleri Kongresi 2013*, (pp. 1)., Ankara, Türkiye. TMMOB. https://obs.hkmo.org.tr/show-media/resimler/ekler/63a02238c63dda1_ek.pdf.

Park, H., Kim, K., & Lee, D. (2019). Prediction of severe drought area based on random forest: Using satellite image and topography data. *Water*, *11*(4), 705. https://doi.org/10.3390/w11040705.

Qi, J., Chehbouni, A., Huete, A., Kerr, Y., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*(2), 119-126. https://doi.org/10.1016/0034-4257(94)90134-1.

Rhee, J., & Im, J. (2017). Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. *Agricultural and Forest Meteorology*, *237-238*, 105-122. https://doi.org/10.1016/j.agrformet.2017.02.011.

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics (Basel)*, *11*(9), 1714. https://doi.org/10.3390/diagnostics11091714.

Sahaar, S. A., & Niemann, J. D. (2024). Estimating rootzone soil moisture by fusing multiple remote sensing products with machine learning. *Remote Sensing*, *16*(19), 3699. https://doi.org/10.3390/rs16193699.

Schroeder, W., Oliva, P., Giglio, L., Quayle, B., Lorenz, E., & Morelli, F. (2016). Active fire detection using Landsat-8/OLI data. *Remote Sensing of Environment*, *185*, 210-220. https://doi.org/10.1016/j.rse.2015.08.032.

Schönauer, M., Prinz, R., Väätäinen, K., Astrup, R., Pszenny, D., Lindeman, H., & Jaeger, D. (2022). Spatio-temporal prediction of soil moisture using soil maps, topographic indices and SMAP retrievals. *International Journal of Applied Earth Observation and Geoinformation*, *108*, 102730. https://doi.org/10.1016/j.jag.2022.102730.

Senay, G. B., Friedrichs, M., Singh, R. K., & Velpuri, N. M. (2016). Evaluating Landsat 8 evapotranspiration for water use mapping in the Colorado River Basin. *Remote Sensing of Environment*, *185*, 171-185. https://doi.org/10.1016/j.rse.2015.12.043.

Sentinel Hub (2025). *Moisture index*. https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/msi/. Accessed 2025, January 20.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting* [Conference proceedings]. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, volume 28, (pp. 802 – 810)., Montreal, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

Sobrino, J. A., Jiménez-Muñoz, J. C., & Paolini, L. (2004). Land surface temperature retrieval from Landsat TM 5. *Remote Sensing of Environment*, *90*(4), 434-440. https://doi.org/10.1016/j.rse.2004.02.003.

Song, Z., Xia, J., Wang, G., She, D., Hu, C., & Hong, S. (2022). Regionalization of hydrological model parameters using gradient boosting machine. *Hydrology and Earth System Sciences*, *26*(2), 505–524. https://doi.org/10.5194/hess-26-505-2022.

Sutariya, S., Hirapara, A., Meherbanali, M., Tiwari, M. K., Singh, V., & Kalubarme,

M. H. (2021). Soil moisture estimation using Sentinel-1 SAR data and land surface temperature in Panchmahal district, Gujarat state. *International Journal of Environment and Geoinformatics (IJEGEO)*, *8*(1), 065-077. https://doi.org/10.30897/ijegeo.777434.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.

Türkeş, M. (2020). *Climate and drought in Turkey*. In Harmancıoğlu, N. B. and Altınbilek, D. (Eds.), *Water resources of Turkey* (pp. 85–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-11729-0_4.

USGS (2019). *Landsat 8 data users handbook*. (LSDS-1574, Version 5.0) Department of the Interior USGS. https://www.usgs.gov/media/files/landsat-8-data-users-handbook.

USGS (2024). *Earth Explorer*. https://earthexplorer.usgs.gov/. Accessed 2024, December 23.

USGS (2025a). *Landsat surface reflectance-derived spectral indices*. https://www.usgs.gov/landsat-missions/landsat-surface-reflectance-derived-spectral-indices. Accessed 2025, January 20.

USGS (2025b). *Landsat surface temperature*. https://www.usgs.gov/landsat-missions/landsat-surface-temperature. Accessed 2025, January 20.

USGS Earth Resources Observation and Science (EROS) Center (2025). *Shuttle Radar Topography (SRTM) mission 1 arc-second global*. https://doi.org/10.5066/F7PR7TFT. Accessed 2025, January 18.

Wang, F., Qin, Z., Song, C., Tu, L., Karnieli, A., & Zhao, S. (2015). An improved mono-window algorithm for land surface temperature retrieval from Landsat 8 thermal infrared sensor data. *Remote Sensing*, *7*(4), 4268–4289. https://doi.org/10.3390/rs70404268.

Wang, Q., Zhao, L., Wang, M., Wu, J., Zhou, W., Zhang, Q., & Deng, M. (2022). A random forest model for drought: Monitoring and validation for grassland drought based on multi-source remote sensing data. *Remote Sensing*, *14*(19), 4981. https://doi.org/10.3390/rs14194981.

Wang, S., & Fu, G. (2023). Modelling soil moisture using climate data and normalized difference vegetation index based on nine algorithms in alpine grasslands. *Frontiers in Environmental Science*, *11*. https://doi.org/10.3389/fenvs.2023.1130448.

Welikhe, P., Essamuah-Quansah, J., Fall, S., & McElhenney, W. (2017). Estimation of soil moisture percentage using Landsat-based moisture stress index. *Journal of Remote Sensing & GIS*, *06*. https://doi.org/10.4172/2469-4134.1000200.

Wilhite, D. A., & Glantz, M. H. (1985). Understanding the drought phenomenon: The role of definitions. *Drought Mitigation Center Faculty Publications*, *20*. https://doi.org/10.1080/02508068508686328.

Wilhite, D. A., Svoboda, M. D., & Hayes, M. J. (2007). Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water Resources Management*, *21*(5), 763–774. https://doi.org/10.1007/s11269-006-9076-5.

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, *27*(14), 3025–3033. https://doi.org/10.1080/01431160600589179.

Xu, Q., Li, S., Yi, J., & Wang, X. (2024). Modeling and application of drought monitoring with adaptive spatial heterogeneity using eco–geographic zoning: a case study of drought monitoring in Yunnan province, China. *Water*, *16*(17), 2500. https://doi.org/10.3390/w16172500.

Yangın, G. (2019). *XGBoost ve karar ağacı tabanlı algoritmaların diyabet veri setleri üzerine uygulaması*. [Master's Thesis]. Mimar Sinan Fine Arts University.

Yıldırım, T., & Asık, Ş. (2018). Index-based assessment of agricultural drought using

remote sensing in the semi-arid region of western Turkey. *Journal of Agricultural Sciences*, *24*(4), 510–516. https://doi.org/10.15832/ankutbd.337136.

Zarei, A., Mahmoudi, M., & Moghimi, M. (2023). Determining the most appropriate drought index using the random forest algorithm with an emphasis on agricultural drought. *Natural Hazards*, *115*, 923–946. https://doi.org/10.1007/s11069-022-05579-2.

Zeng, Y., Feng, Z., & Xiang, N. (2004). *Assessment of soil moisture using Landsat ETM+ temperature/vegetation index in semiarid environment* [Conference proceedings]. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 6, (pp. 4306–4309)., Anchorage, AK, USA. IEEE. https://doi.org/10.1109/IGARSS.2004.1370089.

Zhang, L., Zeng, Y., Zhuang, R., Szabó, B., Manfreda, S., Han, Q., & Su, Z. (2021). In situ observation-constrained global surface soil moisture using random forest model. *Remote Sensing*, *13*(23), 4893. https://doi.org/10.3390/rs13234893.