

Application of machine learning for estimating reference evapotranspiration and crop yield based
on climatological data

by

Sunil Kumar Bhandari

B.S., Tribhuvan University, 2017

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Carl and Melinda Helwig Department of Biological and Agricultural Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Dr. Vahid Rahmani

Copyright

© Sunil Kumar Bhandari 2021.

Abstract

Accurate estimation of reference evapotranspiration (ET_0) provides information on crop water requirement. Knowing crop water requirement in advance allows water managers to take decision for improving the agricultural production and food security at the regional and global scales. ET_0 and crop yield are complex and nonlinear phenomenon depending on various environmental factors. Due to their complex characteristics, this study implemented various machine learning algorithms to predict ET_0 and crop yield. Performance of the implemented models were evaluated using statistical metrics of coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and relative absolute error (RAE).

The overall question this study aimed to address was: will machine learning algorithms be able to model the complex phenomenon of reference evapotranspiration and crop yield across the central United States? To support this research question, two studies were performed in which machine learning algorithms were developed and applied to estimate reference evapotranspiration and crop yield.

The first part of the study evaluates the performance of four machine learning algorithms of support vector machine (SVM), random forest (RF), artificial neural network (ANN), and extreme learning machine (ELM) to predict daily ET_0 . Four input combinations acquired from minimum and maximum temperatures (T_{min} and T_{max}), net solar radiation (R_n), wind speed (U_2), and relative humidity (H_r) were considered using weather data from 1979-2014 for warm season (May through September) from 39 weather stations across the central United States. The predicted ET_0 values were compared with the ET_0 estimated using the FAO-56 Penman-Monteith model. Results revealed that all implemented models showed good results with $R^2 > 0.952$ when all five parameters of T_{max} , T_{min} , R_n , U_2 , and H_r were used for input variables. When

fewer meteorological parameters, including only $T_{\max}$, R_n , and H_r were used, all models still showed satisfactory results with an R^2 not less than 0.868. ELM showed better performance for all input combinations.

The second part of the study implemented six machine learning models, including SVM, ANN, ELM, RF, multiple linear regression (MLR), and deep neural network (DNN) for predicting corn yield using satellite-based vegetation index and weather data from 2003-2020 for growing season (May through September) across Kansas, Iowa, and Nebraska in the Corn Belt, United States. Normalized difference vegetation index (NDVI) is used as satellite-based vegetation index. Biweekly composite MODIS Aqua 1-km NDVI product were used. The predicted crop yield was compared with the USDA yield statistics. The USDA yield statistics consists of corn production in tons per hectare for each county. All models were trained using “with NDVI” and “without NDVI” input combinations. The “without NDVI” input combination weather data of day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation and “with NDVI” input combination includes all parameters of “without NDVI” input combination plus satellite based NDVI. Results revealed that all implemented models showed R^2 range between 0.351-0.634 when used with “with NDVI” input combination and R^2 range between 0.343-0.591 when used with “without NDVI” input combination. Performance of all models were slightly higher using “with NDVI” input combination as compared to using “without NDVI” input combination. DNN showed the best performance following by ELM. MLR showed the worst performance using both the input combination.

The developed machine learning models showed very high accuracy as compared to the other developed machine learning or empirical models for estimating ET_0 . This study could help

to choose the best predictive machine learning models for estimating ET_0 across the central United States with higher accuracy. The developed machine learning models helps to evaluate the ET_0 with fewer parameters with competitive accuracy as compared to FAO-56 PM model. The accurate estimation of ET_0 helps to quantify the crop water requirement in advance and helps to utilize the limited irrigation availability effectively. This helps to increase the irrigation system efficiency and supports in conservation of water resources and irrigation water management.

Secondly, the implemented machine learning models showed average performance for forecasting crop yield across the part of the Corn Belt in the United States. Our model limits the estimation of crop yield before the harvest due to the utilization of NDVI data from the entire growing season. The predictive model would predict the crop yield at any period before the harvest when utilize with NDVI during different stages of the crop growth. The performance of the models could be increased by adding soil data as an input variable. This study helps to provide information on increasing performance of the models by using NDVI as an input.

Table of Contents

List of Figures	viii
List of Tables	ix
Acknowledgements.....	x
Chapter 1- General Introduction	1
Chapter 2 - Evaluating daily reference evapotranspiration using machine learning algorithms with meteorological data in the central United States	7
Abstract.....	7
Introduction.....	8
Materials and methods	13
Study area and data	13
Machine learning models.....	16
<i>Artificial neural network (ANN)</i>	16
<i>Extreme learning machine (ELM)</i>	17
<i>Support vector machine (SVM)</i>	17
<i>Random forest (RF)</i>	18
Input parameters and model development	18
Performance criteria.....	20
Results and discussion	21
Summary and conclusion.....	32
Chapter 3 - Crop yield forecasting with machine learning in the United States Corn Belt	34
Abstract.....	34
Introduction.....	35
Materials and methods	39
Study area and data	39
Machine learning models.....	43

<i>Multiple linear regression (MLR)</i>	43
<i>Artificial neural network (ANN)</i>	43
<i>Support vector machine (SVM)</i>	44
<i>Random forest (RF)</i>	44
<i>Extreme learning machine (ELM)</i>	44
<i>Deep neural network (DNN)</i>	45
Performance metrics	45
Result and Discussion	46
Summary and conclusion	52
Chapter 4 - Summary and Conclusion	54
Chapter 5 - References	57
Appendix A - Mean values of meteorological data for 39 weather stations with location	76

List of Figures

Figure 2-1. Location of 39 weather stations analyzed across the central United States. The dots show the location of HadISD stations. The dots indicating weather station of the same color belong to the same cluster/group.	15
Figure 2-2. Correlation matrix between different climate parameters impacting ET_0	20
Figure 2-3. R^2 of general, each cluster, and average cluster cases for the four machine learning models using different input combinations.	25
Figure 2-4. RMSE of general, each cluster, and average cluster cases for the four machine learning models using different input combinations.	26
Figure 2-5. Scatter plots of the estimated ET_0 values versus the values calculated using FAO-56 PM model for the general case. Each subplot is for one input combination. The models in the legend are ordered according to their R^2 values, lowest value at the top.	27
Figure 2-6. Computational time (in seconds) for all the proposed models for general case using different input combinations. C1, C2, C3, and C4 indicate input combination one, two, three, and four, respectively.	30
Figure 3-1. Study area covers the three states of Kansas, Nebraska, and Iowa in United States Corn Belt.	40
Figure 3-2. The mean corn yield across the study area from 2003 to 2020. Kansas shows the lowest corn yield and Iowa shows the highest corn yield during the period.	41
Figure 3-3. Performance metrics R^2 and RMSE for six machine learning models using with NDVI and without NDVI input combinations.	48
Figure 3-4. Crop yield estimates by the six machine learning models versus the actual crop yield obtained from the USDA NASS. Each subplot is for particular model using both input combinations.	49

List of Tables

Table 2-1. Performance metrics for the four machine learning models with different input parameters during general and cluster case.....	23
Table 3-1. The annual mean values of climatic parameters of the study area.	42
Table 3-2. The performance metrics for six machine learning models with NDVI and without NDVI input combination during training and testing data.	47

Acknowledgements

Firstly, I would like to thank my major professor Dr. Vahid Rahmani and committee members Dr. Vara Prasad and Dr. Trisha Moore for their guidance and support throughout the research work. Secondly, I would like to thank my family and friends for constant love, support, believe, and encouragement throughout the life journey. Finally, I would like to acknowledge the USDA National Institute of Food and Agriculture Project KS545 for funding this research.

Chapter 1- General Introduction

Machine learning is the subset of Artificial Intelligence. Machine learning is a technique that uses previous experience (training data) to learn and use it to perform a new task. It consists of data in a form of set of examples. Each example is described by a set of attributes, also known as variables or features. A feature can be nominal, binary, ordinal or numeric. The performance of the machine learning models in a specific task is measured using performance metrics that are improved with experience over time. At the end, a machine learning model is used to predict, classify, or cluster new examples (testing data) from the learning experience during the training process. Depending on the learning signal, machine learning task can be classified into two categories, that is, supervised and unsupervised learning. In supervised learning, data are labeled (examples with input and corresponding output) and the objective is to construct a general rule that maps inputs to outputs. In unsupervised learning, however, there is no distinction between training and test sets with data being unlabeled. The learner processes input data with the goal of discovering hidden patterns.

Artificial intelligence is a black box that maps the non-linear relationships between input and output variables and can efficiently improve prediction accuracy (Chia, Huang, Koo, et al., 2020; Fan et al., 2018). Therefore, scientists have put forward the application artificial intelligence for predictive purpose. Due to high performance and the powerful advantage of capturing nonlinear and complex structures, machine learning is widely applicable in the field of marketing for tasks classification; in the field of finance for forecasting; in the field of telecommunications for prediction and tasks classification; and in the field of network analysis for relating tasks. In addition, there are wide application of machine learning in the agricultural production system which includes, water management, crop management, livestock

management, and soil management. Water management includes prediction of evapotranspiration (Feng, Peng, et al., 2017) and daily dew point temperature (K. Mohammadi et al., 2015). Crop management includes yield prediction (Kung et al., 2016), disease detection (Moshou et al., 2004), weed detection (Binch & Fox, 2017), crop quality (Maione et al., 2016), and species recognition (Grinblat et al., 2016). Livestock management includes the estimation of farming parameters to optimize the economic efficiency of production system such as animal welfare (Matthews et al., 2017) and livestock production (Alonso et al., 2013). Soil management includes prediction/identification of agricultural soil properties such as estimation of soil temperature (Nahvi et al., 2016), soil condition (Morellos et al., 2016), and soil moisture (Johann et al., 2016).

Among the wide variety of application of machine learning in agriculture, its utility in modeling various hydrological and biophysical process has increased. Furthermore, due to high predictive ability, application of machine learning models for predicting evapotranspiration have dramatically increased since 2011 (Chia, Huang, Koo, et al., 2020). Allen et al., (1998) defined evapotranspiration (ET) as “the combination of water loss from soil surface by evaporation and from crops by transpiration”. Reference ET is the evapotranspiration rate from a reference surface without water source limitation and is generally denoted as ET_0 . ET is a non-linear and complex phenomenon which depends on integration of several climatological factors. It is difficult to estimate ET_0 using empirical approaches. Several studies have used machine learning algorithms to predict ET_0 (Majhi & Naidu, 2021; Reis et al., 2019; Seifi & Riahi, 2020).

Additionally, crop yield has spatial and temporal nonlinear characteristics and depends on numerous complex factors such as crop genotype, environment, and management practices. It is extremely challenging to predict crop yield. Several studies applied numerous machine learning

algorithms to predict crop yield before the harvest (Kim et al., 2020; Kouadio et al., 2018; Sharifi, 2021).

There exist a linear relationship between marketable yield and crop ET for crops such as corn, potato, sugar beet, and wheat (Letey et al., 2007). Many studies used empirical relationship between crop yield and ET for decades to examine the effects of water stress on potential yield reductions (Doorenbos & Kassam, 1979; Howell, 1990). With more than 60% of water contributed to irrigation of crops, agriculture is considered to be world's largest consumer of fresh water (Wada & Bierkens, 2014). According to the NASS, 2017, United States' irrigated cropland encompasses 18% of all the production area but it yields 50% of crop production. This indicate that irrigation plays a critical role in increasing crop yield. We must further increase irrigation efficiency to produce more food with less water to enhance food security at regional and global scales (Smidt et al., 2016). Despite the use of new cultivars, high yields are not achieved if water is limited. It is necessary to utilize the limited available source of water effectively to fulfil crop water requirement. One strategy to achieve this is through increasing irrigation efficiency based on accurate estimation of ET_0 . Accurate quantification of ET_0 in advance helps to develop strategies to increase crop yield. Predicting crop yield before harvest helps farmers and decision makers for better planning and management.

There are various studies applying different machine learning algorithms at local and global scales to predict ET_0 . During the last two decades, different artificial neural network (ANN) models have been applied for estimating ET_0 such as multiple-layer perceptron neural network (Malik et al., 2018), fuzzy logic (Malik & Kumar, 2015), and the adaptive neuro-fuzzy neural network (Keshtegar et al., 2018). The motivation for utilizing the ANN model is due to its high accuracy, robustness, and big data handling (Kisi et al., 2018; Yuan et al., 2018). Support

vector machine (SVM) is also a powerful approach for classification and regression, which attempts to reduce the least square errors. SVM models for predicting ET_0 includes but not limited to least square support vector machine (D. Chen, 2011), SVM-polynomial (Mehdizadeh et al., 2017), and SVM -radial basis function (Seifi & Riahi, 2020). The extreme learning machine (ELM) is a fast learning technique with high generalization capacity that uses single hidden layer feedforward neural network (Kasun et al., 2013). ELM has been used for predicting ET_0 in various studies (e.g. Abdullah et al., 2015; Heddami et al., 2020; Wu et al., 2021). Tree-based models such as random forest (RF) have been applied in many studies to predict ET_0 due to their large data handling capacity and higher level of accuracy over decision tree algorithms (Fan et al., 2018; Filgueiras et al., 2020; Granata, 2019). Other machine learning models used for estimating ET_0 are relevance vector machine (Bachour et al., 2016), gene expression programming (Mattar, 2018), and multivariate regression spline (Mehdizadeh et al., 2017).

There are various machine learning algorithms used for estimating crop yield at local and regional level. These models include but not limited to RF (Dang et al., 2021; Prasad et al., 2021; Sahu et al., 2017), SVM (Bondre & Mahagaonkar, 2019; Oguntunde et al., 2018; Suresh et al., 2020), ANN (Elavarasan & Vincent, 2021; Hara et al., 2021; Liang et al., 2021), multiple linear regression (Shahhosseini et al., 2021; Zaefizadeh et al., 2011), and LASSO (Shook et al., 2021). According to a literature review performed by Van Klompenburg et al., (2020) application of deep learning in predicting crop yield has been increasing. Deep learning models include convolutional neural network (CNN) (Sun et al., 2019; Terliksiz & Altýlar, 2019), long-short term memory (LSTM) (Sharma et al., 2020), recurrent neural network (RNN) (Agarwal & Tarar, 2021; Sivanandhini & Prakash, 2020). Most applied deep learning algorithms include CNN, LSTM, and DNN.

Machine learning is very good at modeling non-linear and complex structure. Due to this advantage of machine learning, our study hypothesize that the machine learning will be able to model the complex agricultural phenomenon of reference evapotranspiration and crop yield with reasonable accuracies. To test this hypothesis, two research were conducted which are included as chapter 2 and chapter 3 in this study.

Previous research showed that the four machine learning models of ELM, SVM, RF, and ANN performance was better as compared to other models when applied on different geographical area. The accuracy of empirical models using few parameters is low as compared to different machine learning models (Gonzalez del Cerro et al., 2020; Reis et al., 2019). Furthermore, very few studies have investigated the performance of these models for estimating ET_0 in the central United States.

In the central United States, irrigation heavily relies on the High Plains Aquifer for agricultural production which has caused a dramatic drop in the water level of the Aquifer from predevelopment (about 1950) to 2015 and from 2013 to 2015 (V. L. McGuire, 2017). Increased irrigation efficiency by providing more accurate ET_0 can improve water conservation and extend the life of the aquifer. To overcome this problem and help in water resource management and irrigation water planning, therefore, we performed research with following objectives (a) development and comparison of machine learning models of ELM, SVM, RF, and ANN to estimate daily ET_0 with various meteorological parameters in the central United States. (b) To find an optimal climatic parameter that works with all the applied models in prediction with reasonable accuracies. (c) To find out the machine learning model with highest predictive power with optimal climatic parameter across the central United States.

There are various other studies that showed the highest predictive capabilities of machine learning models than other empirical method in estimating crop yield across the different climatic condition. Developing reliable model to predict crop yield in advanced helps farmers and decision makers for better planning and management. Recently, the Food and Agriculture Organization of the United Nations (FAO) reported maize to be largest grain crop produce in the world. Taylor & Koo, (2011) claimed United States to be the largest grain producer and exporter supplying approximately 40% of global annual corn production. Corn is among the world's top three grain crops having substantial importance for the global economy. To overcome the problem of food security and helps in global economy, we performed research with following objectives (a) comparison of different machine learning models to forecast corn yield across the part of the Corn Belt in the United States. (b) To check weather adding NDVI to the weather variable as input increases the performance of the applied models.

Chapter 2 - Evaluating daily reference evapotranspiration using machine learning algorithms with meteorological data in the central United States

Abstract

Due to an increase in drought events throughout the world, the incessant need to increase agricultural production requires a careful assessment of irrigation water planning. Accurate prediction of reference evapotranspiration (ET_0) plays a critical role in irrigation water consumption and conservation of infinite water resources. This study evaluates the performance of four machine learning models, including the extreme learning machine (ELM), support vector machine (SVM), random forest (RF), and artificial neural network (ANN) for estimating daily ET_0 using various meteorological data for input variables. Four input combinations acquired from minimum and maximum temperatures (T_{min} and T_{max}), net solar radiation (R_n), wind speed (U_2), and relative humidity (H_r) were considered using meteorological data from 1979-2014 for warm seasons (May through September) from 39 weather stations across the central United States. Performance of the models was evaluated by comparing their outputs with the ET_0 estimated using the FAO-56 Penman-Monteith model. Four different statistical metrics of coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and relative absolute error (RAE) were applied to assess performance of the models. Two strategies were used to develop the models and measure their efficiency: (i) general case: models are developed using data from all the weather stations (ii) cluster case: weather stations are clustered/grouped with similar climatic characteristics, using K-means algorithms, and models were developed for each group using data from weather stations in that group. All the models

were trained and tested using 1979-2004 and 2005-2014, data respectively. Results in the first strategy revealed that all implemented models showed good results with an $R^2 > 0.961$ when five parameters of $T_{\max}$, $T_{\min}$, R_n , U_2 , and H_r were used for input variables. ELM performed best with an $R^2 = 0.976$ and $RMSE = 0.363 \text{ mm d}^{-1}$. When fewer meteorological parameters, including only $T_{\max}$, R_n , and H_r were used, all models still showed satisfactory results with an R^2 not less than 0.903. ELM performed best with an $R^2 = 0.917$ and $RMSE = 0.678 \text{ mm d}^{-1}$. The Second strategy also revealed similar result with ELM showing the best performance using all the four-input combination and RF being the least accurate. In addition, ELM showed the smallest computational costs. Considering prediction accuracy, reliability, and computational cost of the studied models, ELM can be recommended when used with complete and fewer meteorological parameters, respectively, across the central United States. These models can be utilized for estimating ET_0 in other parts of the world depending on their prediction accuracy, particularly when fewer climate parameters are available.

Keywords: extreme learning machine, support vector machine, random forest, artificial neural network, irrigation, artificial intelligence

Introduction

A progressive increase in the world population and drought conditions has led to the necessity of increased agricultural production (Tavakol, A., & Rahmani, V., 2020a, 2020b). Irrigation plays a significant role in supporting this production. Evapotranspiration (ET) is a key contributor to hydrological water balance and water resources planning and decision making for irrigation management. An accurate estimation of ET is essential for irrigation practices, and hence, fulfillment of crop water requirements. Allen et al. (1998) defined ET as “the combination of water loss from the soil surface by evaporation and from crops by transpiration.”

Reference ET is the evapotranspiration rate from a reference surface without water source limitation and is generally denoted as ET_0 . ET_0 is a climatic parameter and depends on climatic variables. It can be measured by either a direct method such as using a lysimeter (Anapalli et al., 2016; Shrestha & Shukla, 2015) or estimated with indirect methods such as using meteorological data. Nevertheless, the measurement with a lysimeter is expensive, time consuming, and requires precise planned experiments (Chia, Huang, & Koo, 2020; Moorhead et al., 2019). Indirect methods, which vary from empirical relationships to complex functions based on different meteorological data, have been widely used for estimating ET_0 .

The Penman-Monteith model, which incorporates both thermodynamic and aerodynamic aspects of the ET phenomenon, has shown to be relatively more accurate than other existing empirical approaches for estimating ET_0 (Fan et al., 2018; Timm et al., 2014). The Food and Agriculture Organization of the United States (FAO) has recommended FAO-56 PM as the universal model among various models for estimating ET_0 for any climatic conditions across the world (Richard G Allen et al., 1998). However, the FAO-56 model requires a large number of input variables such as minimum temperature (T_{min}), maximum temperature (T_{max}), wind speed (U_2), relative humidity (H_r), and net solar radiation (R_n) (Chia, Huang, & Koo, 2020; Fan et al., 2018). These climate parameters are not always available for specific times and geographic areas, which is a major drawback of the model.

ET is a non-linear and complex phenomenon that depends on integration of several climatological factors and is difficult to estimate using empirical approaches. Scientists have put forward the application of artificial intelligence (AI) for estimating ET. AI is a black box that maps the non-linear relationships between input and output variables and can efficiently improve prediction accuracy (Chia, Huang, Koo, et al., 2020; Fan et al., 2018). In the past few decades,

different AI techniques, including machine learning methods, have been increasingly utilized to estimate different hydrological and renewable energy variables such as drought (Kousari et al., 2017), solar radiation (Voyant et al., 2017), streamflow (Parisouj et al., 2020), evaporation (Keshtegar et al., 2016), and rainfall (Abrahart et al., 2012). Due to high performance and the powerful advantage of capturing nonlinear and complex structures, applications of machine learning models for predicting ET have dramatically increased since 2011 (Chia, Huang, Koo, et al., 2020).

Various machine learning techniques have been applied both on local and global scales to predict ET_0 , which includes but is not limited to (1) support vector machine (SVM) (Chia, Huang, & Koo, 2020; B. Mohammadi & Mehdizadeh, 2020; Seifi & Riahi, 2020); (2) extreme learning machine (ELM) (Chia et al., 2021; Reis et al., 2019; Wu et al., 2019); (3) artificial neural networks (ANN) (Adnan et al., 2017; Althoff et al., 2018; Genaidy, 2020; Majhi & Naidu, 2021) such as multiple layer perceptron (MLP), back propagation with different learning rates, radial basis function, and momentum; (4) tree-based models such as random forest (RF) (Althoff et al., 2018; Fan et al., 2018; Filgueiras et al., 2020; Granata, 2019; S. Wang et al., 2019) and MP5 regression tree and bagging (Granata, 2019); and (5) other machine learning models such as the adaptive neuro fuzzy inference system (Dou & Yang, 2018; Gonzalez del Cerro et al., 2020), relevance vector machine (RVM) (Deo et al., 2016), multivariate adaptive regression spline (Deo et al., 2016), and gene expression programming (S. Wang et al., 2019).

Yang et al., (2006) applied SVM, multiple regression and MLP with back propagation to predict ET in non-forest sites across the continental United States using remotely sensed data. They found SVM performed better than these two models. Moghaddamnia et al., (2008) found good ET_0 estimates using SVM with the radial basis function as the kernel function for

Chahnimeh reservoirs in southeastern Iran. In terms of prediction accuracy and stability, SVM showed similar results to the ELM model when applied to predict daily ET_0 across different climates in China (Fan et al., 2018). The SVM model was superior compared to ANN and RVM for predicting crop ET and crop coefficients in south Florida (Shrestha & Shukla, 2015). ANN and SVM models were developed for different clusters of weather stations and showed better performance when compared with various empirical models for estimating ET_0 across the Brazil (Ferreira et al., 2019b). Similarly, deep learning models were compared with classical machine learning models using clustered weather stations for predicting ET_0 across the Northeast plain, China (Z. Chen et al., 2020). Dou & Yang, (2018) employed the ELM model for predicting daily ET across different terrestrial ecosystems in Germany and Sweden and found that ELM can be recommended over ANN and SVM for forest and wetland ecosystems, respectively. In another study, ELM showed highest prediction accuracy when compared with ANN, SVM, and Hargreaves methods for modelling weekly ET in the Thar Desert in India (Patil & Deka, 2016). Using temperature-based data as input variables for ELM, ET_0 was accurately estimated across the semi-arid region of the Verde River Basin in southern Brazil (Reis et al., 2019). In another study in the same region, ANN performed better than the original Hargreaves Samani method for predicting ET_0 (Reis et al., 2019). The performance of ANN for estimating ET_0 was better in grassland and cropland ecosystems as compared to ELM and ANFIS models in the United States and Belgium (Dou & Yang, 2018). RF was compared with SVM, bagging, and MP5 regression tree to predict ET in a site in central Florida with pastures of Bahia grass (Granata, 2019). They found the tree-based model, MP5 regression tree, has the best predictive capabilities. Filgueiras et al., (2020) compared the RF model with five regression models for predicting actual ET and soil water content using remote sensing data across the tropical climate of Brazil. Their results

indicated that RF accurately estimated soil water content. Chia, Huang, and Koo, (2020) evaluated the performance of SVM using different input variables for estimating daily ET_0 across the tropical climate of Malaysia. It was found that SVM, using only solar radiation and mean temperature as input variables, performed the best.

Among the mentioned machine learning models, ELM, SVM, RF, and ANN exhibited better predicting accuracies compared to other models. These four models have been increasingly used for predicting ET_0 in different locations worldwide. However, the application of machine learning techniques for estimating ET_0 is very limited in the central United States. This region has been experiencing increase in dry (Stone & Rahmani, 2020; Tavakol et al., 2020b) and windy (Tavakol et al., 2020c) conditions and heat waves (Tavakol et al., 2020a). The region is predominantly agricultural including rangeland; natural prairie; and irrigated and rain-fed farming of maize, soybean, sorghum, alfalfa, cotton, and sugar beets (Mutiibwa & Irmak, 2013). Most of the irrigation water in this region is provided from the High Plains Aquifer and other infinite water resources.

Previous studies showed that the four machine learning models of ELM, SVM, RF, and ANN performed better compared to other models in different parts of the world. However, the performance of machine learning models depends on the geographical area in addition to the availability of input variables (Dou & Yang, 2018; Yaseen et al., 2020). Furthermore, very limited studies have investigated the performance of these models for predicting ET_0 in the central United States. The objective of this study includes (a) development and comparison of machine learning models of ELM, SVM, RF, and ANN to estimate daily ET_0 with various meteorological parameters in the central United States. (b) To find an optimal climatic parameter that works with all the applied models in prediction with reasonable accuracies. (c) To find out

the machine learning model with highest predictive power with optimal climatic parameter across the central United States. Information for 39 weather stations in warm season from 1979-2014 were analyzed. Two strategies were used to develop the models and measure their efficiency: (i) general case: models were developed using data from all the weather stations and (ii) cluster case: weather stations were clustered with similar climatic characteristics, using K-means algorithms, and models were developed for each group using data from weather stations in that group.

Materials and methods

Study area and data

This study covers the central United States, which also has been designated as the Great Plains (Figure 2-1). The region consists of 10 states of Montana, Wyoming, North Dakota, South Dakota, Colorado, Nebraska, Kansas, New Mexico, Oklahoma, and Texas. Topographical characteristics of the area vary from the lowest elevation of 35 meters in the southern coastline in Texas in the south, to flat-to-rolling plains with the highest elevation of 1,884 meters in Colorado in the west. The long-term (1979-2014) warm season daily average values of $T_{\min}$ and $T_{\max}$ range from 11 °C in the north to 22 °C in the south, and 26 °C in the north to 33 °C in the south, respectively. Humid and warm air masses flowing from the Gulf of Mexico, and cold air fronts from the Rocky mountains in the northwest and Canada in the north, dominate climatic conditions of the region (Irmak et al., 2012). The long-term daily average U_2 ranges from 2.8 ms^{-1} in Montana in the northwest to 5.9 ms^{-1} in Kansas in the east. Daily average H_r varies from 36 % in Colorado in the west to 76 % in Texas in the south. Daily average R_n varies from 226 Wm^{-2} in North Dakota in the north to 292 Wm^{-2} in Colorado in the west. Similarly, long-term daily

average ET_0 varies from 6 mm d^{-1} in North Dakota in the north to 9.6 mm d^{-1} in Texas in the south.

Data for $T_{\min}$ and $T_{\max}$, U_2 , and H_r were obtained from the HadISD (www.metoffice.gov.uk/hadobs/hadisd/), a sub-daily, station-based and quality controlled dataset based on the integrated surface database (ISD) from the National Oceanic and Atmospheric Administration's National Centre for Environment Information (NOAA/NCEI, formerly the National Climatic Data Centre-NCDC) (Dunn et al., 2016; Tavakol et al., 2020b). The R_n dataset was obtained from the NCDC's Solar and Meteorological Surface Observation Network (SAMSON, archived by NCDC, Ashville, NC), National Solar Radiation Database (NSRDB, archived by the National Renewable Energy Lab), and NCDC's Integrated Surface Hourly Database (NCDC-ISH) (www.ncdc.noaa.gov/isd). Daily values of all five parameters of $T_{\min}$, $T_{\max}$, U_2 , H_r , and R_n from 1979-2014 for warm seasons (May through September) were obtained from 39 weather stations across the region. For each station, when any of the five parameters were missing, values for that day were omitted from the analysis.

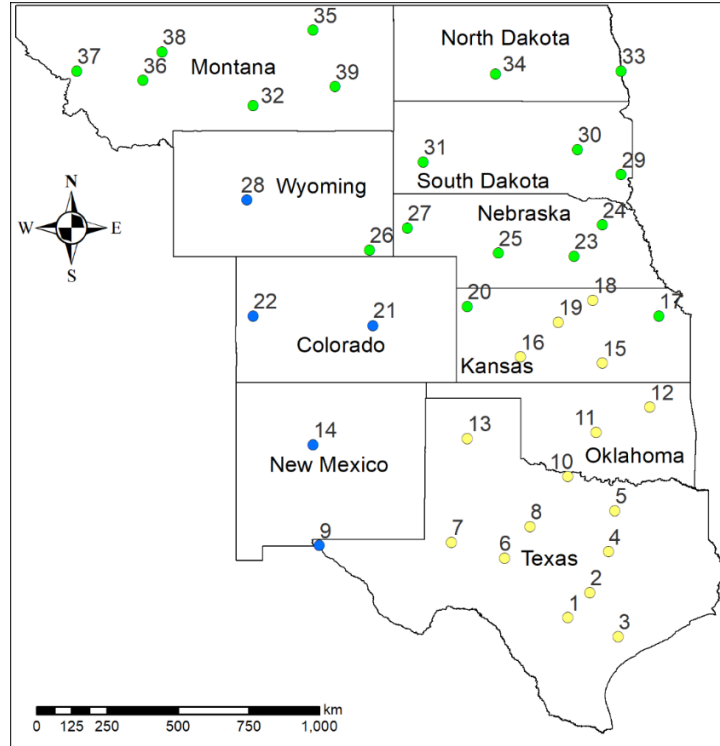


Figure 2-1. Location of 39 weather stations analyzed across the central United States. The dots show the location of HadISD stations. The dots indicating weather station of the same color belong to the same cluster/group.

FAO-56 Penman-Monteith model

The FAO-56 PM method recommended by the FAO (Richard G Allen et al., 1998) was used to estimate daily ET_0 . This method is universally accepted for predicting ET_0 in different climates (Bellido-Jiménez et al., 2021; Moratiel et al., 2019; Wen et al., 2015).

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (1)$$

where ET_0 is the reference evapotranspiration (mm d^{-1}), R_n is the net radiation at the crop surface ($\text{MJ m}^{-2} \text{d}^{-1}$), G is the soil heat flux density ($\text{MJ m}^{-2} \text{d}^{-1}$), T is the mean daily air temperature at 2 meters height ($^{\circ}\text{C}$), u_2 is the wind speed at 2 meters height (ms^{-1}), e_s is the saturation vapor

pressure (kPa), e_a is the actual vapor pressure (kPa), $e_s - e_a$ is the saturation vapor pressure deficit (kPa), Δ is the slope of the vapor pressure curve ($\text{kPa}^\circ\text{C}^{-1}$), and γ is the psychrometric constant ($\text{kPa}^\circ\text{C}^{-1}$). More details on can be accessed from the FAO-56 publication (Richard G Allen et al., 1998). The estimated ET_0 using eq(1) acts as an output variable to the proposed, supervised machine learning models. In addition, performance of the models was evaluated by comparing the estimated ET_0 from the models with ET_0 values estimated using eq(1).

Machine learning models

Artificial neural network (ANN)

ANN model was first introduced by Rosenblatt (1958) as a computational method. It is one of the most classically supervised learning models that is analogous to the biological neural network, which has been applied in several studies to estimate ET_0 (Ferreira et al., 2019a; Maroufpoor et al., 2020; Yassin et al., 2016). ANN is a parallel distributed processing system that consists of an input layer, output layer, and one or more hidden layers (Sudheer et al., 2003). Each layer consists of a node or neuron and every node in one layer is connected to every other node in the next layer. Each node takes the weighted sum of its inputs and passes it through the non-linear activation function. The output of one node becomes the input of another node in the next layer. The signal flows from left to right and the final output is calculated by performing the procedure for all nodes. The processing element consists of two parts. The first aggregates the weighted input and the second is a non-linear filter, usually called the activation function. This process is described as a forward pass in which given an input and weights, the output is computed. To train our model to learn the weights, the backward pass from right to left was performed and propagated the error to every individual node using back propagation. Advantages of using back propagation are to derive optimal weights and minimize errors

between the desired target and the modeled values, i.e., loss function. The predictive ability of the ANN model is strongly dependent on its variety of internal parameters and functions such as network topology structure, activation functions, and data partitioning method (Maier et al., 2010). In this study, selection of internal parameters was based on trial and error.

Extreme learning machine (ELM)

The ELM was initially proposed by Huang et al. (2006). This model has been applied in several studies to estimate ET_0 (e.g. Gocic et al., 2016; Reis et al., 2019; Zhu et al., 2020). ELM is a single-hidden-layer, feed-forward neural network, unlike the conventional back-propagation-based ANN model. It consists of three layers: an input layer, a hidden layer, and an output layer. In the ELM, the number of neurons does not need to be obtained using a trial-and-error procedure as in the ANN. Moreover, adjustment of input weights and biases terms are not required as they are independent of training data. The speed of ELM is extremely fast with better generalization capacity and no problems of overfitting and local minima (Guang Bin Huang et al., 2006).

Support vector machine (SVM)

SVMs are supervised learning algorithms developed by Cortes and Vapnik (1995) and deal with classification, regression, and pattern-recognition problems. The SVM works by transforming the lower-dimensional input data into higher-dimensional feature space and maps their relationship with the output data. The transformation of input vector is controlled by various kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. The type of selected kernel function depends on the field of application. RBF is one of the most commonly deployed kernel functions for predicting the ET_0 (Fan et al., 2018; Granata, 2019; Mehdizadeh et al., 2017; Yin et al., 2017). The RBF kernel function was used in this study.

Random forest (RF)

The RF model proposed by Breiman (2001) is one of the most effective ensemble algorithms. RF has been used in many studies to estimate the ET_0 (e.g., Fan et al., 2018; Feng et al., 2017; Wang et al., 2019). The RF model consists of random binary trees and employs a subset of the training dataset through the bootstrapping technique. The random subset of the training dataset is sampled from the raw dataset and the out-of-bag dataset (the dataset not included in the model development) is applied to evaluate general problems. During the process of RF development, groups of input variables are randomly selected to split at each node. Suppose there are M input variables such that at each node, m variables ($m \ll M$) are randomly selected out of the M . The best split on these m is used to divide the node. The value of m is kept constant throughout the growing of the forest.

Input parameters and model development

To evaluate performance of the proposed machine learning models for estimating ET_0 using different meteorological data, four different input combinations were established using T_{max} , T_{min} , U_2 , H_r and R_n (Althoff et al., 2018; Chia et al., 2021; Fan et al., 2018; Feng, Cui, et al., 2017). Input combination one included all five climatic parameters. The remaining input combinations were created by dropping a parameter having a lower correlation with ET_0 (Figure 2-2). Particularly, input combination two excludes U_2 ; input combination three excludes U_2 and T_{min} ; and input combination four excludes H_r , T_{min} , and U_2 . Each of these four different input combinations of climatic variables act as one input set to the proposed models as shown in Table 2-1. The weather stations were grouped according to the similar climatic characteristics using K-means algorithms. For the general case, models were developed using data from all weather stations and in the cluster case, models were developed for each cluster using data from weather

stations in that cluster. All models were trained using data from 1979 - 2004 and tested using data from 2005 - 2014 data. The performance metrics of the models from each cluster were averaged and presented in Table 2-1 along with the general case. Similarly, the values of R^2 and RMSE of each model for general, each cluster, and average cluster case is shown in the Figures 2-3 and 2-4, respectively. Daily mean values of $T_{\min}$, $T_{\max}$, U_2 , H_r , R_n and ET_0 from 1979 - 2014 were available for each weather station. To reduce the impact of different scales, these features were standardized according to eq(2) and are used to group all weather stations by K-means algorithm.

$$x_{ni} = \frac{x_i - \mu}{\sigma} \quad (2)$$

where x_{ni} is the standardized value, x_i is the observed value, μ is the mean and σ is the standard deviation. K-means is an unsupervised algorithm which aims objects of the same group to have high similarity and objects of different groups to have low similarity (Zhou & Zhao, 2013). It partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The clustering process is based on the proximity between objects measured by Euclidean distance. Previous studies used Silhouette coefficient to determine best number of groups and Euclidean to determine distance measure (Z. Chen et al., 2020; Ferreira et al., 2019a). This study also adopted same technique and obtained three clusters with a best Silhouette coefficient of 0.381.

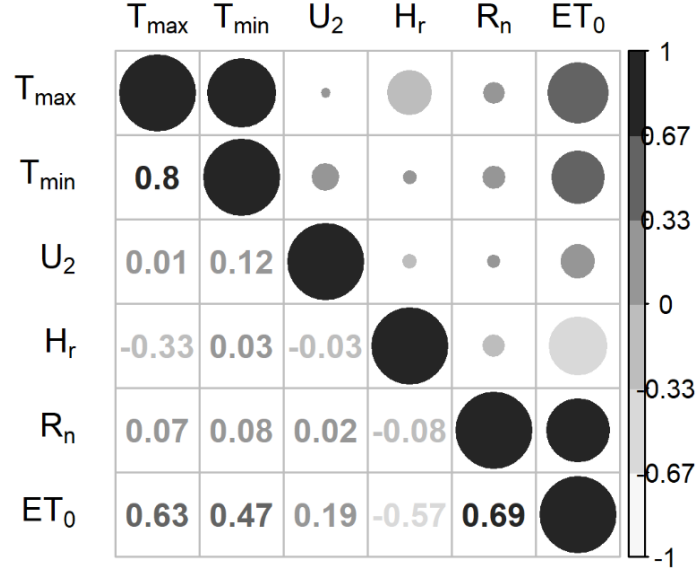


Figure 2-2. Correlation matrix between different climate parameters impacting ET_0 .

Performance criteria

The performance of machine learning forecasting models was assessed by means of four different statistical metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and relative absolute error (RAE) using the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (5)$$

$$RAE = \frac{\sum_{i=1}^n |y_i - x_i|}{\sum_{i=1}^n |\bar{x} - x_i|} \quad (6)$$

where x_i , y_i , $\bar{x}$, and n are the measured value, model estimated value, mean of measured value, and total number of observations, respectively.

R^2 ranges between 0 and 1 and values closer to 1 indicate better performance. RMSE, MAE, and RAE range from 0 to infinity and lower values (closer to zero) of these metrics indicate better performance for a model. These statistical indicators have been widely used when comparing different models for predicting ET_0 (Fan et al., 2018; Landaras et al., 2008).

Results and discussion

Predicted ET_0 values using four machine learning models were compared with Penman-Monteith estimates. Penman-Monteith values were estimated using the FAO-56 PM method (Richard G Allen et al., 1998). Four statistical metrics of RMSE, MAE, RAE, and R^2 were used to measure the goodness-of-fit of each model. Each ELM, SVM, RF, and ANN model was performed with the four different input combinations. Each input combination consisted of different climate variables selected based on their correlation with ET_0 (Figure 2-2). The results shown in Table 2-1 represent the performance of the proposed models for the general and cluster cases.

For the general case, the result revealed that models performed differently using various input combinations (Table 2-1). Applying input combination one for predicting daily ET_0 , all four models, with R^2 values of 0.961 or higher, performed best compared to using other input combinations. ELM performed best with the highest R^2 (0.976) and lowest error values of RMSE (0.363 mmd^{-1}), MAE (0.256 mmd^{-1}), and RAE (13.753 mmd^{-1}) followed by RF, SVM, and ANN (Table 1). However, results obtained by all four models were very close to each other suggesting that when using $T_{\max}$, $T_{\min}$, U_2 , H_r , and R_n for the input variables, all four machine learning methods can predict daily ET_0 with high accuracy. Using input combination two, ELM showed the best performance ($R^2 = 0.936$, RMSE = 0.597 mmd^{-1} , MAE = 0.403 mmd^{-1} , and RAE = 21.677 mmd^{-1}) followed by ANN, RF, and SVM. There was a 4-6 % decrease in values of R^2

among all models using input combination two versus one. This small difference indicates that performance of machine learning models without using U_2 among the input parameters is still satisfactory for estimating daily ET_0 . All models indicated similar performances applying input combination two. Using input combination three, ELM showed the best performance ($R^2 = 0.917$, $RMSE = 0.678 \text{ mmd}^{-1}$, $MAE = 0.485 \text{ mmd}^{-1}$, and $RAE = 26.096 \text{ mmd}^{-1}$) followed by ANN, SVM, and RF. Like using input combinations one and two, all models showed similar performances using input combination three. R^2 values decreased by less than 2% for input combination three versus two. With R^2 values greater than 0.903 for all models using input combination three, the models were effective in predicting daily ET_0 . This suggests that without including U_2 and $T_{\min}$ among the input variables, all four machine learning methods can provide daily ET_0 estimates with high accuracy. Finally, using input combination four, ELM showed the best performance ($R^2 = 0.799$) followed by ANN, SVM, and RF. There was a 12-15 % decrease in R^2 values for all models using input combination four versus three. This large decrease indicates that models not including U_2 , $T_{\min}$, and H_r for input parameters, are less likely to estimate ET_0 accurately as performance drops considerably.

Table 2-1. Performance metrics for the four machine learning models with different input parameters during general and cluster case.

Input combinations/ Parameters	Machine Learning Models	general case				cluster case			
		R ²	RMSE	MAE	RAE	R ²	RMSE	MAE	RAE
			(mmd ⁻¹)	(mmd ⁻¹)	(%)		(mmd ⁻¹)	(mmd ⁻¹)	(%)
Input combination 1 (T _{max} , T _{min} , U ₂ , H _r , R _n)	ELM	0.976	0.363	0.256	13.753	0.977	0.338	0.237	13.310
	SVM	0.971	0.397	0.303	16.295	0.954	0.471	0.404	22.907
	RF	0.975	0.387	0.308	16.544	0.952	0.458	0.389	22.255
	ANN	0.961	0.464	0.325	17.472	0.954	0.466	0.399	22.633
Input combination 2 (T _{max} , T _{min} , H _r , R _n)	ELM	0.936	0.597	0.403	21.677	0.933	0.585	0.405	22.638
	SVM	0.910	0.702	0.547	29.422	0.905	0.678	0.539	30.494
	RF	0.911	0.686	0.536	28.801	0.898	0.689	0.546	30.932
	ANN	0.916	0.676	0.520	27.957	0.905	0.673	0.530	30.061
Input combination 3 (T _{max} , H _r , R _n)	ELM	0.917	0.678	0.485	26.096	0.912	0.664	0.478	26.736
	SVM	0.904	0.723	0.547	29.402	0.875	0.713	0.558	32.107
	RF	0.903	0.724	0.549	29.533	0.868	0.789	0.620	35.061
	ANN	0.905	0.722	0.546	29.356	0.880	0.760	0.596	33.747
Input combination 4 (T _{max} , R _n)	ELM	0.799	1.079	0.807	43.375	0.803	1.018	0.757	42.463
	SVM	0.763	1.138	0.859	46.209	0.727	1.143	0.877	49.661
	RF	0.761	1.130	0.854	45.930	0.703	1.164	0.891	50.335
	ANN	0.788	1.079	0.815	43.838	0.728	1.139	0.876	49.632

ELM: extreme learning machine, SVM: support vector machine, RF: random forest, ANN: artificial neural network, T_{max}: maximum temperature, T_{min}: minimum temperature, U₂: wind speed, H_r: relative humidity, R_n: net solar radiation

In cluster case, there are three groups consisting of 5, 16, and 18 weather stations. The performance of the models from all the clusters are averaged as shown in the Table 2-1.

Applying input combination one for predicting daily ET_0 , all four models, with R^2 values of 0.952 or higher, performed best compared to using other input combinations. ELM performed best with the highest R^2 (0.977) and lowest error values of RMSE (0.338 mmd^{-1}), MAE (0.237 mmd^{-1}), and RAE (13.310 mmd^{-1}) followed by ANN, SVM, and RF (Table 1). ELM performed slightly better in cluster case but ANN, SVM, and RF showed better performance in general case. However, all of the models performed was similar in both general and cluster cases. Using input combination two, ELM showed the best performance ($R^2 = 0.933$, RMSE = 0.585 mmd^{-1} , MAE = 0.405 mmd^{-1} , and RAE = 22.638 mmd^{-1}) followed by ANN, SVM, and RF. All models show slightly lower performance compared to the general case. Using input combination three, ELM showed the best performance ($R^2 = 0.912$, RMSE = 0.664 mmd^{-1} , MAE = 0.478 mmd^{-1} , and RAE = 26.736 mmd^{-1}) followed by ANN, SVM, and RF. The performance of all models was slightly weaker as compared to the general case. Finally, using input combination four, ELM showed the best performance ($R^2 = 0.803$) followed by ANN, SVM, and RF. ELM performed slightly better in cluster case but ANN, SVM, and RF showed better performance in general case. There was an 11-19 % decrease in R^2 values for all models using input combination four versus three. This large decrease indicates that models with not including U_2 , $T_{\min}$, and H_r for input parameters are less likely to estimate ET_0 accurately as performance drops considerably even in the cluster case.

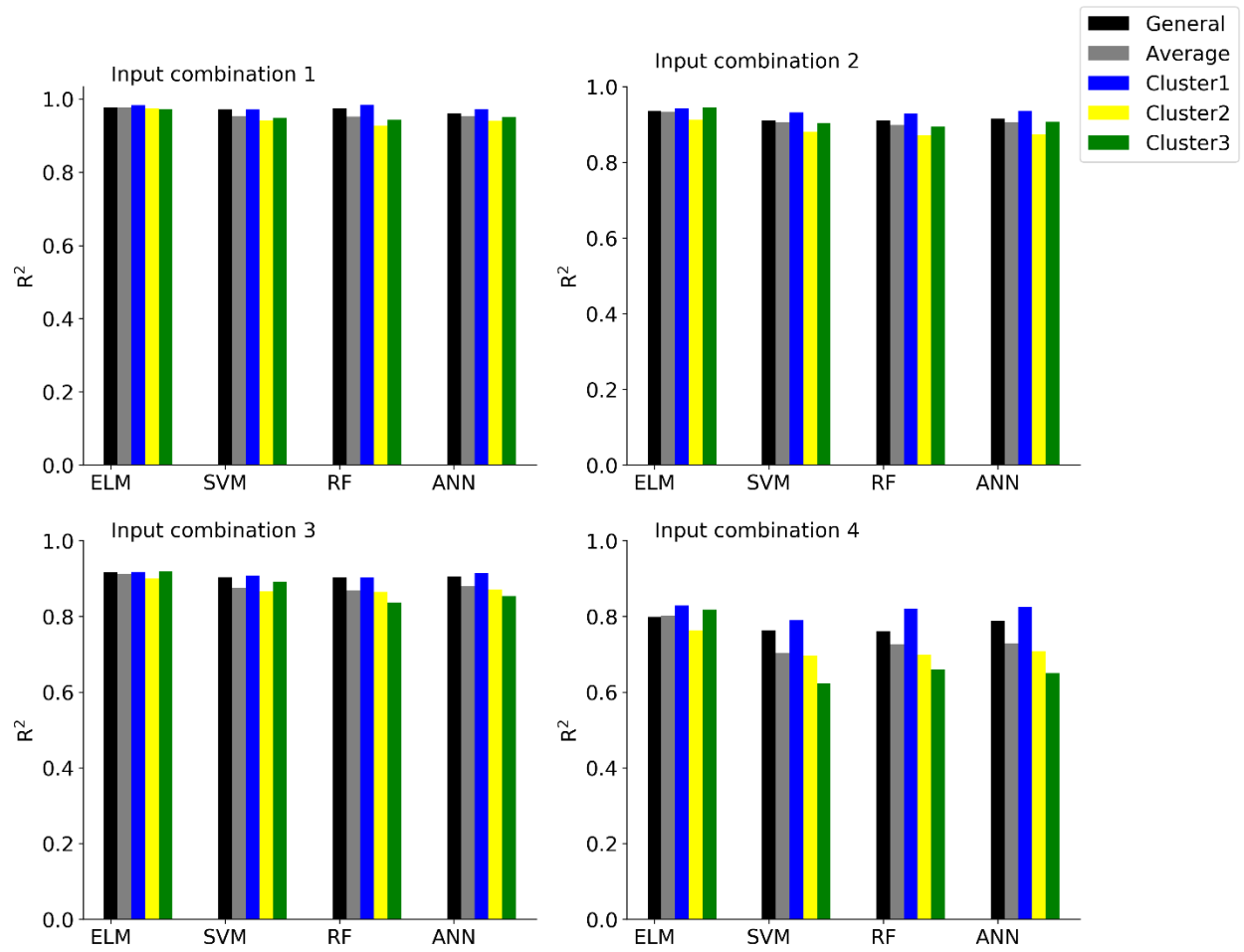


Figure 2-3. R^2 of general, each cluster, and average cluster cases for the four machine learning models using different input combinations.

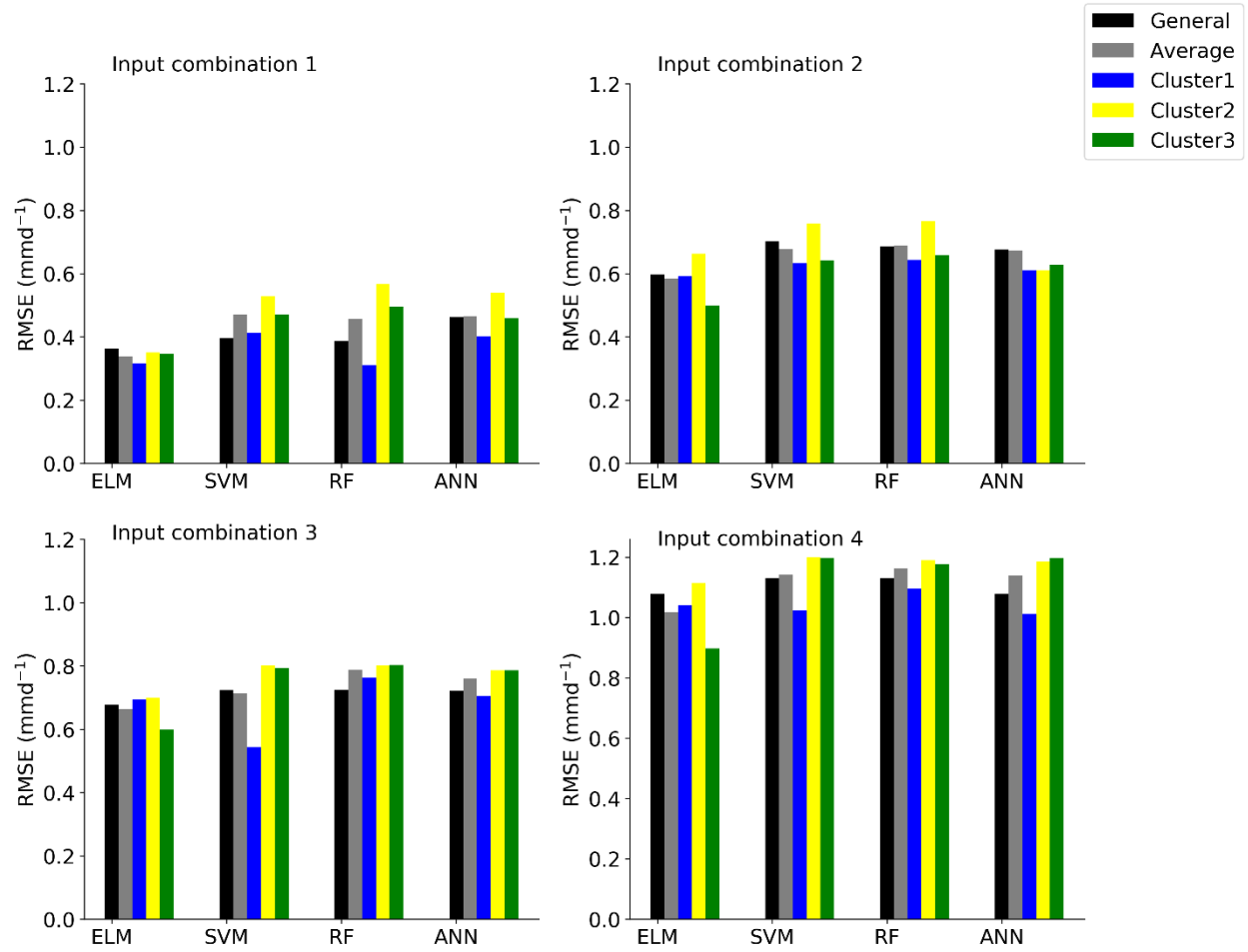


Figure 2-4. RMSE of general, each cluster, and average cluster cases for the four machine learning models using different input combinations.

ELM indicated best predictive accuracies using all four input combinations for both cases. Using input combination one, ANN performed weak in general case and RF performed weak in cluster case. Using input combination two, SVM and RF exhibited lower performance in general and cluster cases, respectively. However, using input combination three and four, RF performed worst in both cases. This indicates that performance of RF model is less reliable as compared to all other models. Figures 2-3 and 2-4 show the performance metrics R^2 and RMSE for the proposed model using four different input combinations for general case, cluster case (average result of individual clusters), and individual clusters, respectively. Most of the models

performed slightly better in cluster 1 as compared to the general case. Cluster 2 and 3 performed slightly worse than the general case. However, overall, most of the models showed slightly better performance in the general case as compared to the cluster case (Figures 2-3 and 2-4). Changes in R^2 values between general and cluster case ranged from 0.1% - 7.6% for all models and input combinations. This indicates that there is no considerable difference in the performance of the models between general and cluster cases which indicates the reliability of the models in different climatic conditions.

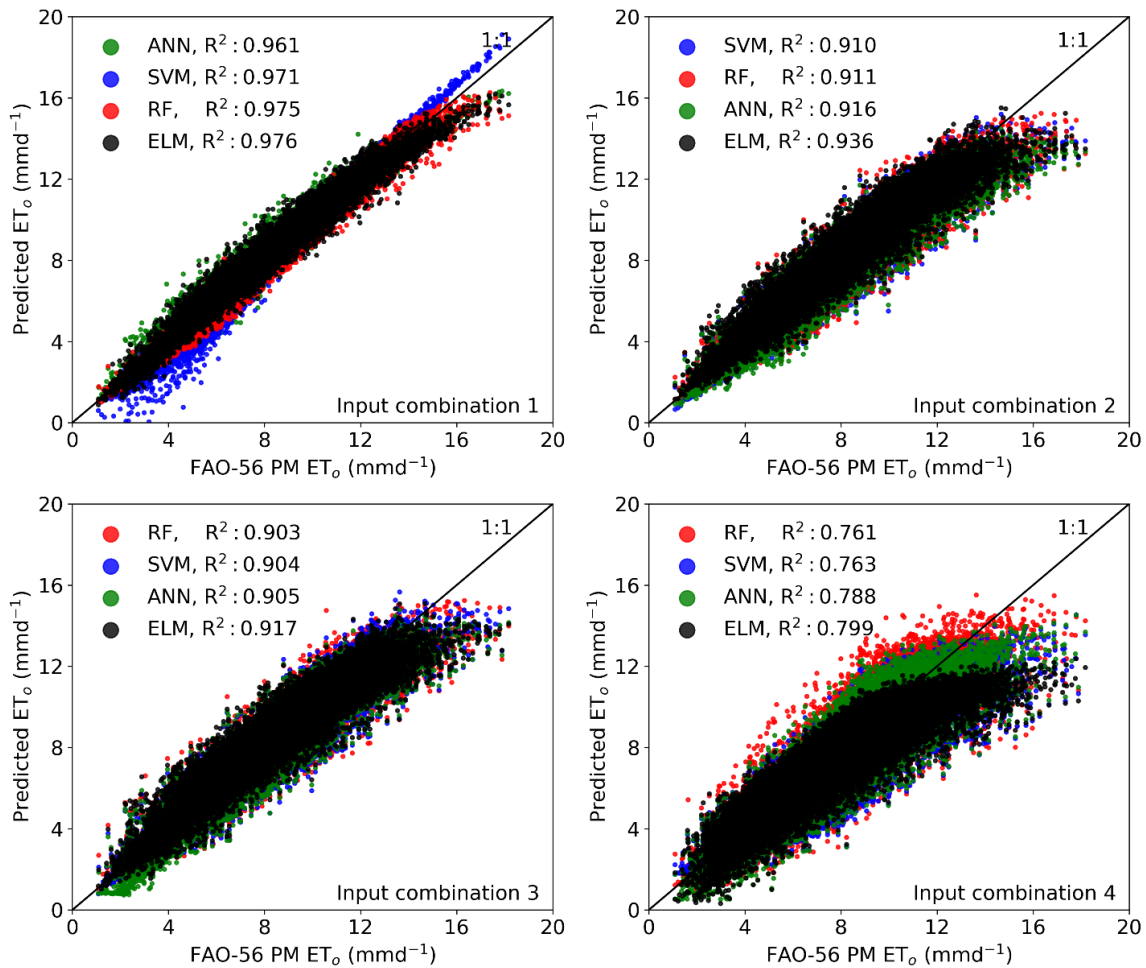


Figure 2-5. Scatter plots of the estimated ET_0 values versus the values calculated using FAO-56 PM model for the general case. Each subplot is for one input combination. The models in the legend are ordered according to their R^2 values, lowest value at the top.

Daily ET_0 values calculated by the FAO-56 PM model were plotted against daily ET_0 values predicted by the four machine learning models for the testing period in the general case (Figure 2-5). Each subplot included the values estimated by all four machine learning models using one of the input combinations. The dispersion degree of the data points in the scatter plots was lowest for input combination one, followed by input combinations two, three, and four, respectively.

Our findings indicate that all the adopted machine learning models are powerful tools for predicting daily ET_0 . Number and type of input variables implemented in the models played an important role in determining accuracies of these models. Suleiman and Hoogenboom (2007) used the same five input variables as input combination one to estimate daily ET_0 with FAO-56 PM and Priestley-Taylor models across Georgia. They found that FAO-56 PM ET_0 values were very close to measured ET_0 . Results are generally in accordance with previous studies, which had concluded that more input variables improve model performance (Althoff et al., 2018; Chia et al., 2021; Fan et al., 2018; Tabari et al., 2012). ELM shows the highest accuracy for predicting weekly ET for the various input combinations from T_{min} , T_{max} , solar radiation, and ET_0 in the Thar desert of India (Patil & Deka, 2016). SVM, ELM, and other tree-based models showed better performance than other input combinations when used with these five parameters to predict daily ET_0 in the temperate continental, mountain plateau, and temperate monsoon zones of China (Fan et al., 2018). ELM was best ET_0 estimator over ANN and SVM in forest and wetland ecosystems in Germany and Sweden, respectively (Dou & Yang, 2018). Wavelet ELM model showed highest accuracy for predicting daily ET_0 when used with different input combinations from temperature, solar radiation, relative humidity, and wind speed data across Turkey (Kisi & Alizamir, 2018). Feng et al. (2017) found that random forest and generalized

regression neural network showed good performance for predicting daily ET_0 in the humid region of southwest China using the same input variables as in input combination one in this study. Results of this study showed similar performance metrics for all four machine learning models when U_2 data was eliminated from input variables in input combination two. This is particularly critical when spatial and/or temporal U_2 data is not available. It should be noted that the Penman-Monteith approach suggests assuming a 2 ms^{-1} fixed value for wind speed when direct measurements are not available, recommending the lesser impact of this variable on the estimated values for ET_0 . Córdova et al., (2015) further supported the assumption of Penman-Monteith and showed the negligible difference in performance with and without using U_2 among input variables for estimating daily ET_0 . Granata (2019) carried out a study in central Florida with pastures of Bahia grass to predict actual ET. The author concluded that RF and SVM models showed satisfactory performance when R_n , mean H_r , and mean temperature were included in the input variables, and U_2 was excluded. ELM showed highest accuracy as compared to Hargreaves Samani method to predict daily ET_0 using only temperature data in Verde River Basin, Southern Brazil (Reis et al., 2019). ELM with genetic algorithm optimization technique shows reliable performance when implemented with only temperature data across the different climates of China (Gong et al., 2021). When both U_2 and $T_{\min}$ were eliminated from input variables (input combination three), all four machine learning methods demonstrated high performance metrics. The performances were very close to when only U_2 was eliminated for input combination two, suggesting that in addition to U_2 , excluding $T_{\min}$ does not considerably decrease the prediction power. This result agreed with the study conducted by Chia et al., 2021 in which the difference in predictive accuracies measured by R^2 was very optimal using input combination two versus three. This is particularly important when temporal and/or spatial

information is not available for $T_{\min}$. Finally, when H_r is eliminated from input variables (input combination four), model performances dropped remarkably. Althoff et al., (2018) applied five models to estimate daily ET_0 and compared them with the FAO-56 PM ET_0 using different input combinations across the Minas Gerais in Brazil. They found that dropping U_2 among input variables of mean temperature, R_n , H_r , and U_2 , the performance of SVM and RF models decreased considerably as compared to using all input variables. Shiri et al. (2014) applied several models including ANN and SVM to estimate ET_0 in hyper-arid, semi-arid, and humid regions in Iran. SVM and ANN provided small RMSE values ranging from 0.359 to 0.704 mmd^{-1} when $T_{\min}$, $T_{\max}$, mean temperature, R_n , and H_r were used for input variables. Generally, model performance decreased when only temperature data and R_n were used for input variables with RMSE ranging from 0.470 to 0.804 mmd^{-1} . Using only temperature and solar radiation data to predict daily ET_0 across the Poyang Lake basin of South China, ELM model performance was better than RF (Wu et al., 2021).

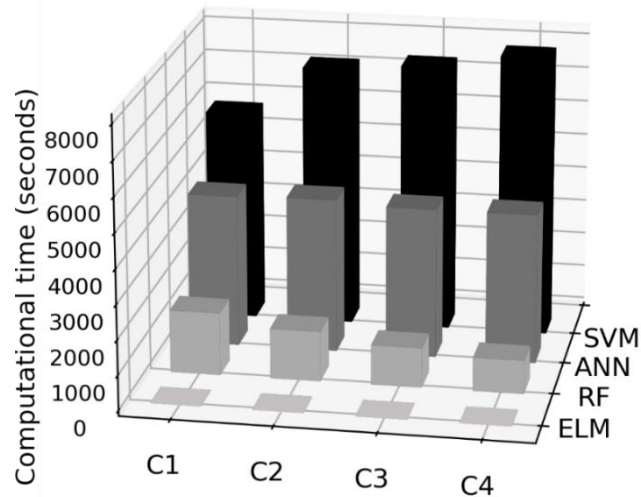


Figure 2-6. Computational time (in seconds) for all the proposed models for general case using different input combinations. C1, C2, C3, and C4 indicate input combination one, two, three, and four, respectively.

Training and testing machine learning models have been implemented using dataset from all weather stations within the study area (Ferreira et al., 2019b; Reis et al., 2019). ANN was best and ELM was worst for grassland ecosystem but ELM was best and ANN was worst for forest ecosystem (Dou & Yang, 2018). To overcome this problem, our study used K-means method to cluster the weather stations into three groups having similar climatic conditions. In each cluster, all proposed models were trained and tested using weather stations from the same cluster. The results indicated that the proposed models had no considerable difference in their accuracies for the cluster case as compared to the model performance in general case. Ferreira et al., 2019b found no considerable difference in performance between general case and cluster case for predicting daily ET_0 by SVM, ANN, and empirical models in Brazil. The present study used large amount of data that are spatially and temporally variant, which shows greater reliability of the proposed models.

Computational costs/time of all proposed models were measured using different input combinations (Figure 2-6). SVM and RF were implemented using a grid search to find the best combination of parameters among 45 and 144 different parameter combinations, respectively. Computational costs of SVM were higher compared to other models for all of the four input combinations. Computational costs of SVM increased when number of input parameters dropped from five for input combination one to two for input combination four. RF showed less computational costs compared to SVM and ANN. Computational costs of SVM models were 3.4-8.4 times greater than RF. ANN was implemented using 5,000 iterations to obtain higher values of R^2 and lower values of prediction errors. Computational costs were high for ANN compared to ELM and RF, and lower than SVM. ELM indicated the least computational costs for all of the input combinations. In a study by Fan et al. (2018), computational costs of SVM

were up to 1.4 times higher than the computational costs of RF and ELM for estimating daily ET_0 across different climatic zones in China. In another study, Hassan et al. (2017) indicated the computational costs of SVM were up to 0.7-19.5 times higher than the computational costs of RF and four other models for estimating global solar radiation in five different countries. In terms of computational costs, ELM showed the greatest predicting speed following by RF, ANN, and SVM. All computation analysis was performed with the following computing features: Intel(R) Xeon(R) CPU E3-1245 v5 @ 3.50GHz, and 16 GB Ram memory.

Summary and conclusion

An accurate prediction of ET_0 is crucial for water resource management and irrigation water planning. Application of four machine learning algorithms, including ELM, SVM, RF, and ANN, were evaluated for estimating daily ET_0 . The four machine learning models indicated high prediction accuracy using different input combinations. Analyses was conducted for 39 weather stations across the central United States from 1979–2014 during the warm season (May through September). Using four input combinations from meteorological data (T_{max} , T_{min} , R_n , H_r , and U_2) as input variables, all models performed well, with ELM showing the best performance for each combination. Models using input combination two (T_{max} , T_{min} , R_n , and H_r) showed similar results compared to models using input combination three (T_{max} , R_n , and H_r). This indicates lack of wind speed and minimum temperature information does not impact model efficiency considerably. Performance of the models decreased dramatically when only T_{max} and R_n were used for input variables, suggesting that using only T_{max} and R_n may not provide good estimates of ET_0 . In addition, RF, SVM, and ANN showed the lowest performance and highest computational costs as compared to the ELM. Results recommended that ELM is reliable, accurate, and less expensive models when used with both complete and limited input data for predicting daily ET_0 .

across the central United States. In the central United States, irrigation heavily relies on the High Plains Aquifer for agricultural production which has caused a dramatic drop in the water level of the Aquifer from predevelopment (about 1950) to 2015 and from 2013 to 2015 (V. L. McGuire, 2017). Increased irrigation efficiency by providing more accurate ET_0 can improve water conservation and extend the life of the aquifer (Deines et al., 2021). In addition, the accurate estimation of ET_0 helps to quantify the crop water requirement in advance and therefore, help farmers to utilize the limited irrigation water effectively which increases the irrigation efficiency. This work also helps farmers to manage the irrigation water requirement and conserve the water resources. The future study includes evaluating the performance of machine learning models at different time scales (hourly or monthly) or in different climatic conditions across the world or using different technique of forming input combination rather than based on correlation coefficient.

Chapter 3 - Crop yield forecasting with machine learning in the United States Corn Belt

Abstract

Accurate predictions of crop yield are crucial for enhancing food security at the regional and global scales. The United States supplies approximately 40% of the global annual corn production and claimed to be the world largest grain producer and exporter. Developing reliable model to accurately predict crop yield helps to assure food security by guiding the farmers, planning food storage and transport, and by helping policymakers focus on the most vulnerable communities. Yield prediction can help prevent famine, support local economy, and improve sustainable agricultural practices. Due to spatiotemporal nonlinear characteristics of crop yield, this study implemented six machine learning models, including support vector machine (SVM), artificial neural network (ANN), extreme learning machine (ELM), random forest (RF), multiple linear regression (MLR), and deep neural network (DNN) for predicting corn yield using satellite-based vegetation index and weather data from 2003-2020 for growing season (May through September). Analyses were conducted across the three states in midwestern United States (Kansas with 105 counties, Nebraska with 93 counties, and Iowa with 99 counties) that are part of the Corn Belt. All the models were trained using “with NDVI” and “without NDVI” input combination. The “without NDVI” input combination includes weather data of day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation. The “with NDVI” input combination includes all the parameters of “without NDVI” input combination plus satellite based NDVI. Three different statistical metrics of coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were applied to assess the performance of the models. Results revealed that all the

implemented models showed R^2 range between 0.351-0.634 when used with the “with NDVI” input combination and R^2 range between 0.343-0.591 when used with the “without NDVI” input combination. Results indicated that performance of all models was slightly higher using “with NDVI” input combination as compared to the “without NDVI” input combination. There is approximately up to 23% increase in predictive capabilities of all applied models when NDVI information was used. The DNN model showed the highest performance following by ELM. The MLR indicated the least performance using both input combinations. The results could outbreak the limitation of ELM model to predict corn yield across the Corn Belt in the United States. However, performance of the applied machine learning models can be further improved using data related to soil water in addition to weather and satellite-based information for predicting corn yield.

Keywords: crop yield, food security, corn production, artificial intelligence, machine learning, deep learning, normalized difference vegetation index

Introduction

It is estimated that nearly 795 million people in the world lack sufficient food (S. McGuire, 2015) and there will be two billion more people to feed by 2050 (Dodds & Bartram, 2016). Improving food security is a primary objective in the 2030 Agenda for Sustainable Development of the United Nations (Cf, 2015). A central challenge of ensuring food security is possible only through the development of reliable crop yield model, which helps to accurately predict crop yield before harvest. Developing reliable model to accurately predict crop yield helps to assure food security by guiding the farmers, planning food storage and transport, and by helping policymakers focus on the most vulnerable communities. Yield prediction can help prevent famine, support local economy, and improve sustainable agricultural practices. Also,

farmers and other decision makers can utilize the precise yield prediction for better planning and decision-making (Paswan & Begum, 2013). However, crop yield prediction is extremely challenging due to its dependency on numerous complex factors such as crop genotype, environment, and management practices. Additionally, genotype and environmental factors interaction makes the yield prediction more challenging.

Recently, artificial intelligence (AI) branch machine learning and deep learning techniques have been applied for crop yield prediction, which include but is not limited to regression tree (RT), random forest (RF), multiple linear regression (MLR), support vector machine (SVM), artificial neural network (ANN), and deep neural network (DNN). Machine learning approaches treat output (crop yield) as an implicit function of the input variables, which would be a highly non-linear and complex function. Due to its powerful advantage of capturing nonlinear and complex structures, applications of machine learning models for prediction purposes have been dramatically increased.

Regression models have been widely used to predict crop yield in various studies (Abdipour et al., 2019; Shahhosseini et al., 2021; Zaefizadeh et al., 2011). The use of regression model is limited especially in complex cases like extreme data values, non-linear relationships, and multicollinearity (Zaefizadeh et al., 2011). Due to this limitation, regression model is used as benchmark in various previous studies for the comparison with other models. Simulation method has potential to specify relevant factors affecting yield. However, simulation requires consideration of biophysical inputs that demands estimation instead of measurement. This method requires expert knowledge and user skills to calibrate and more data storage constraints (Shahhosseini et al., 2021). RF shows competitive accuracies when compared with deep learning and neural network models in predicting crop yield in many of the studies (Agarwal & Tarar,

2021; Khaki et al., 2020; Kim & Lee, 2016). RF outperformed MLR when applied for predicting wheat, maize, and potato yield using weather and biophysical data at global and regional scale in the United States (Jeong et al., 2016). SVM shows best prediction accuracies when used with different input combination as compared to deep learning model to predict crop yield (corn) using remote sensing and weather data in Illinois (Kuwata & Shibasaki, 2015). Cai et al., 2019 used weather and satellite data to predict wheat yield in Australia using SVM, RF, and ANN. Five machine learning algorithms along with SVM was compared for predicting crop yield (corn) under extreme weather condition using satellite and meteorological data (Kim et al., 2020). SVM shows similar accuracies as compared to the different deep learning models when used to predict different crop yield (maize, rice, soybean) using climate and soil data in India (Agarwal & Tarar, 2021). ANN outperformed MLR model for estimating winter wheat yields based on remotely sensed and climate data (Jiang et al., 2004). ANN was very effective in predicting corn and soybean yield in Maryland using different input variables of soil and rainfall data (Kaul et al., 2005). ANN was applied to predict maize yield in Jilin China using climate and fertilizer data as predictors (Matsumura et al., 2015). Kross et al., 2018 evaluated the predicting power of ANN for corn and soybean yield using remotely sensed data in eastern Ontario, Canada. Adisa et al., 2019 applied ANN using different input variable combination to predict maize production in different province of South Africa. ANN was compared with DNN, and other machine learning models to predict corn yield using satellite and climate data in midwestern United States (Kim et al., 2020). ANN was applied and compared with RT model to predict crop yield using satellite and climate data in southern Iran (Sharifi, 2021). The performance of 18 different extreme learning machine (ELM)-based models with different combinations of the predictor variables based on soil fertility properties was evaluated and

compared with MLR and RF to predict coffee yield in Lam Dong Province, Vietnam (Kouadio et al., 2018). This study found that ELM performance was better than MLR and RF. Similarly, Maimaitiyiming et al., 2019 found that weighted regularized ELM using combination of hyperbolic tangent and rectified linear unit activation function outperformed RF, MLR, and partial least squares regression in predicting grapevine berry yield in Missouri, United States. In another study by Taherei Ghazvinei et al., 2018, it was found that ELM outperformed ANN and genetic programming model when used for predicting sugarcane growth based on meteorological data in Iran. However, the application of ELM to predict corn yield is very limited across the globe. The application of machine/deep learning in predicting crop yield has been widely increased since 2013 (Van Klompenburg et al., 2020). There are various studies applying deep learning algorithms especially when used with remote sensing data (Kim et al., 2020; Kim & Lee, 2016; A. X. Wang et al., 2018). convolutional neural network (CNN) – a deep learning algorithms was applied to predict wheat and barley using remote sensing data (Nevavuori et al., 2019). Shook et al., 2021 applied recurrent neural network (RNN) to predict soybean yield using genotype and weather variables in Canada and United States. They found that RNN performance was better than least absolute shrinkage and selection operator (LASSO) and SVM. Sun et al., 2019 applied CNN, long-short term memory (LSTM), and CNN-LSTM model to predict soybean yield using crop growth and environmental variables in the continental United States. There are other various machine learning algorithms applied to predict crop yield such as bayesian neural network (Johnson et al., 2016), multivariate adaptive regression spline (Kim et al., 2020), naïve bayes (M. Paul et al., 2015; Sujatha & Isakki, 2016). Among the mentioned algorithms, ANN, CNN, LSTM, and DNN are most widely used machine learning algorithms in crop yield prediction (Van Klompenburg et al., 2020). In addition, Chlingaryan et al., 2018 also

found that ANN, SVM, RF, and deep learning (CNN, RNN, LSTM) are widely applicable for predicting crop yield in precision agriculture (Chlingaryan et al., 2018).

In this study, we focused on building multiple AI models to predict corn yield in the midwestern United States. Six different AI models included SVM, ANN, ELM, RF, MLR, and DNN. For this, we prepared the input data, including satellite-based vegetation index and weather data for May through September (the dominant growth period) during 2003-2020. The objective of this study includes (a) comparison of different machine learning models to forecast corn yield across the part of the Corn Belt in the United States. (b) To check weather adding NDVI to the weather variable as input increases the performance of the applied models.

Materials and methods

Study area and data

Corn is among the world's top three grain crops having substantial importance for the global economy. In 2018, the Food and Agriculture Organization of the United Nations (FAO) reported maize to be largest grain crop produce in the world. Taylor & Koo, 2011 claimed United States to be the largest grain producer and exporter supplying approximately 40% of global annual corn production. This study covers three states in midwestern United States (Kansas with 105 counties, Nebraska with 93 counties, and Iowa with 99 counties) that are part of the Corn Belt where most of the corn is produced in the United States (Figure 3-1).

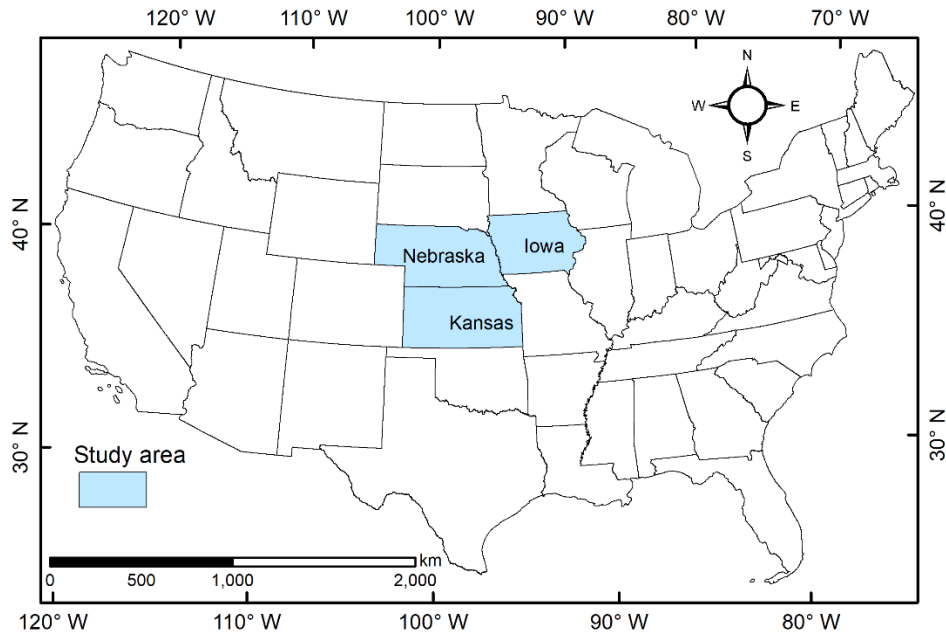


Figure 3-1. Study area covers the three states of Kansas, Nebraska, and Iowa in United States Corn Belt.

Normalized difference vegetation index (NDVI), weather, and corn yield information were used from 2003-2020 during warm season (May through September). The USGS EROS Data Center provides Moderate Resolution Imaging Spectroradiometer (MODIS) data for Earth environmental monitoring. The NDVI information was acquired from the MODIS database as indicators of vegetation vitality, biomass, and photosynthesis. Biweekly composite MODIS Aqua 1-km NDVI product were used. The weather data included daily record of seven weather variables including day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation. The weather data was acquired from Daymet (Thornton et al., 2014). The spatial resolution of the weather data is 1 km. The National Agricultural Statistics Service operated by the USDA provides annual corn yield statistics at the county level available at <https://www.nass.usda.gov/>. The corn yield data in bushel per acre were converted to a ton per hectare. Figure 3-2 shows the time-series plot of annual corn yield in the

study area including Kansas, Nebraska, and Iowa. The mean annual climatic condition of the study area is shown in Table 3-1.



Figure 3-2. The mean corn yield across the study area from 2003 to 2020. Kansas shows the lowest corn yield and Iowa shows the highest corn yield during the period.

Table 3-1. The annual mean values of climatic parameters of the study area.

Year	Corn yield (ton/ha)	Day length (sec)	Snow water equivalent (kg/m ²)	Vapor pressure (pa)	Solar radiation (w/m ²)	Maximum temperature (°C)	Minimum temperature (°C)	Precipitation (mm)	NDVI
2003	8.9	49,986	0.00	1,582	393	28	14	402	5,897
2004	10.8	49,983	0.00	1,560	386	27	14	506	6,127
2005	9.9	49,989	0.00	1,704	389	29	15	477	6,123
2006	9.2	49,993	0.00	1,623	385	28	14	420	5,949
2007	10.1	50,000	0.00	1,763	381	28	15	559	6,173
2008	10.4	50,095	0.01	1,583	379	27	13	610	6,392
2009	11.2	50,039	0.00	1,536	381	26	13	520	6,401
2010	9.6	50,017	0.00	1,744	375	28	15	667	6,257
2011	9.3	50,022	0.00	1,660	385	28	14	458	6,112
2012	7.8	50,011	0.00	1,513	406	30	14	257	5,573
2013	10.0	50,030	0.01	1,649	383	28	14	453	6,116
2014	11.2	50,038	0.00	1,585	380	27	14	613	6,407
2015	11.6	50,068	0.00	1,706	368	27	15	634	6,462
2016	11.7	50,056	0.00	1,753	378	28	15	584	6,556
2017	11.4	50,040	0.00	1,643	387	28	14	465	6,390
2018	11.5	50,056	0.00	1,783	366	28	16	657	6,512
2019	11.5	50,061	0.00	1,748	365	27	15	715	6,544
2020	10.9	49,999	0.00	1,633	382	28	14	433	6,212

Machine learning models

Multiple linear regression (MLR)

MLR is a widely used statistical method, which attempts to model the correlation between dependent and independent variables in a dataset. MLR has been applied for estimating yield estimation for different crops (e.g. Gonzalez-Sanchez et al., 2014; Zaefizadeh et al., 2011). In this method, a dependent variable is linearly related to multiple independent variables and can be written as

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m \quad (1)$$

where $\hat{y}$ is the predicted dependent variable, $x_{1,2,\dots,m}$ are the independent variables and $a_{0,1,\dots,m}$ are the coefficients of the independent variables. These parameters are determined during training phase.

Artificial neural network (ANN)

ANN was first introduced by Rosenblat, 1958 as a computational method. It is a machine learning algorithm that is analogous to the biological neural network, which has been applied in several studies to estimate crop yield (e.g. Bose et al., 2016; Dahikar & Rode, 2014). ANN is a parallel distributed processing system that contains input, hidden, and output layers. Each neuron in each layer is calculated from the previous layer by first calculating a weighted sum of all of its inputs and then applying an activation function. The hyper-parameters such as the number of hidden layers, the number of neurons in each layer, an activation function, learning rate, momentum, and other factors are tuned for each model. The predictive ability of ANN model is highly dependent on its variety of internal parameters, network topology structure, activation

functions, and data partitioning method (Maier et al., 2010). In this study, selection of hyperparameters was based on trial-and-error procedure.

Support vector machine (SVM)

SVM was first developed by Cortes & Vapnik, 1995 which deals with the classification and regression problems. SVM have been implanted for regression model especially for crop yield prediction (Brdar et al., 2011; R. Li et al., 2009). In the SVM regression, the lower-dimensional input data are transformed into a high-dimensional feature space using kernel function. The type of kernel function deployed depends on the relationship between dependent and independent variables. Hyperparameters that need tuning include parameters of kernel function, error margin and other factors. The various kernel functions include linear, polynomial, radial basis function (RBF) and sigmoid. The RBF kernel function was deployed in this study.

Random forest (RF)

The RF model was proposed by (Breiman, 2001) is a tree-based ensemble method built on the concept of bagging. Bagging reduces the prediction variance by averaging the predictions made by sampling with replacement. RF has been applied in many studies to predict various crop yield (Everingham et al., 2016; Jeong et al., 2016). RF randomly choose a random number of features and construct a regression tree with them and repeat this procedure many times and finally average all the prediction made by all regression trees. Hence, RF addresses both bias and variance and is proved to be powerful (Cutler et al., 2007).

Extreme learning machine (ELM)

The ELM was initially proposed by (G.-B. Huang et al., 2006). It is a single hidden layer feedforward neural network with randomly initialized input weights and biases. It consists of three layers: an input layer, a hidden layer, and an output layer. Due to its easy implementation,

fast learning, better generalization capacity, and no problems of overfitting and local minima (G.-B. Huang et al., 2006), ELM has gained more popularity compared to other machine learning methods (Gao Huang et al., 2015). However, very few studies has explored the potential of ELM in crop yield prediction (Maimaitiyiming et al., 2019; Taherei Ghazvinei et al., 2018).

Deep neural network (DNN)

DNN is a new approach in the field of machine learning research, which was originally proposed by Hinton & Salakhutdinov, 2006. DNN regression network is similar to ANN with multiple hidden layers between input and output layers which can model complex non-linear relationships. A fully connected feed-forward DNN has been successfully applied to various regression problems for crop yield prediction especially corn (Khaki & Wang, 2019; Kim et al., 2020). The generic ANN has a problem of local minima and overfitting and hence DNN overcame these problems in more intensive network where backward and forward optimizations are combined.

Performance metrics

In this paper, coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used to evaluate the performance of machine learning forecasting model. The mathematical formulas are:

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (4)$$

where x_i , y_i , $\bar{x}$, and n represents the measured value, model predicted value, mean of measured value, and total number of observations, respectively.

R^2 value ranges between 0 and 1. Similarly, RMSE and MAE values range from 0 to infinity. R^2 close to 1 and RMSE and MAE close to zero indicate the better performance of a model. These statistical metrics have been widely used in previous studies for comparison of different machine learning model to predict crop yield (Dang et al., 2021; Sagan et al., 2021; Shahhosseini et al., 2021).

Result and Discussion

Six different machine learning models were used including MLR, ANN, SVM, RF, ELM, and DNN for predicting corn yield. The estimated corn yields for growing season (May-September) were compared with the USDA yield statistics. Two different input combinations were formed based on weather and satellite data. First input combination includes weather data and NDVI and is called “with NDVI” input combination hereafter. Second input combination includes all the input as in the “with NDVI” input combination excluding the NDVI information. This combination is denoted as the “without NDVI” input combination hereafter (Table 3-2). Each model is implemented using both input combinations and their performance is evaluated using three statistical metrics of RMSE, MAE, and R^2 . To avoid spatial and temporal effect, 75% of randomly selected samples were used in the training phase and 25% are used for testing phase.

For the testing data, results indicated that models performed slightly different using the two input combinations (Table 3-2). Applying “with NDVI” input combination for predicting corn yield, all six models, with R^2 values ranging from 0.351-0.634, performed better compared to using “without NDVI” input combination with R^2 values ranging from 0.343-0.591. DNN

performed the best with the highest R^2 (0.634) and lowest error values of RMSE (1.547 ton/ha) and MAE (1.191 ton/ha) following by ELM, RF, SVM, ANN, and MLR. Similarly, applying “without NDVI” input combination, DNN performed the best with the highest R^2 (0.591), and lowest error values of RMSE (1.635 ton/ha) and MAE (1.258 ton/ha). There is approximately up to 23% increase in values of R^2 for all models using “with NDVI” versus “without NDVI” input combination. This large increase in R^2 values indicates that performance of machine learning models improved using NDVI information along with the weather data.

Table 3-2. The performance metrics for six machine learning models with NDVI and without NDVI input combination during training and testing data.

Input combinations	Machine learning models	Training			Testing		
		R^2	RMSE (ton/ha)	MAE (ton/ha)	R^2	RMSE (ton/ha)	MAE (ton/ha)
With NDVI	MLR	0.385	2.013	1.539	0.351	2.061	1.592
	ANN	0.560	1.703	1.330	0.499	1.810	1.421
	SVM	0.596	1.631	1.242	0.520	1.771	1.354
	RF	0.614	1.595	1.237	0.546	1.724	1.349
	ELM	0.669	1.476	1.133	0.632	1.558	1.212
	DNN	0.720	1.358	1.053	0.634	1.547	1.191
Without NDVI	MLR	0.384	2.015	1.542	0.343	2.073	1.602
	ANN	0.502	1.812	1.383	0.404	1.975	1.527
	SVM	0.562	1.699	1.289	0.501	1.807	1.377
	RF	0.611	1.602	1.244	0.542	1.730	1.364
	ELM	0.643	1.534	1.196	0.571	1.696	1.292
	DNN	0.699	1.409	1.094	0.591	1.635	1.258

Without NDVI input combination includes weather data of day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation. The with NDVI input combination includes all the parameters of “without NDVI” input combination plus satellite based NDVI

For training data, the six models performed slightly different using the two input combinations (Table 3-2). Applying “with NDVI” input combination for predicting corn yield, all six models, with R^2 values ranging from 0.385-0.720, performed best compared to using “without NDVI” input combination with R^2 values ranging from 0.384-0.699. DNN indicated the best performance with the highest R^2 (0.720) and lowest error values of RMSE (1.358 ton/ha) and MAE (1.053 ton/ha) following by ELM, RF, SVM, ANN, and MLR. Similarly, applying “without NDVI” input combination, DNN performed the best with the highest R^2 (0.699), and lowest error values of RMSE (1.409 ton/ha) and MAE (1.094 ton/ha). There is approximately up to 11% increase in values of R^2 for all models using “with NDVI” versus “without NDVI” input combination. This large increase in R^2 values indicates that performance of machine learning models would increase using NDVI information along with the weather data.

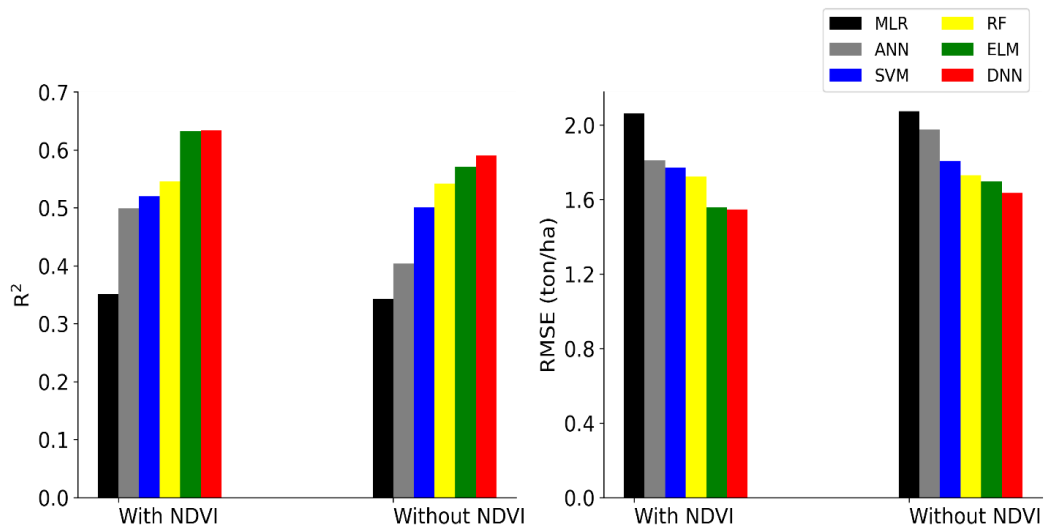


Figure 3-3. Performance metrics R^2 and RMSE for six machine learning models using with NDVI and without NDVI input combinations.

With both training and testing data, DNN and MLR indicated the best and least predictive accuracies, respectively for both input combinations. All models exhibited lower decrease in their performance for predicting corn yield between training and testing data (Figure 3-3).

Changes in R^2 , RMSE, and MAE for training and testing data ranged from 5-19%, 2-16%, and 3-14% for all models and input combinations. Results indicate good stability of all models by showing small changes for prediction errors between training and testing data.

Predicted crop yield values using six machine learning models were compared to USDA actual yield values for the testing data (Figure 3-4). Each subplot included the values estimated by each model using “with NDVI” and “without NDVI” input combination. The dispersion degree of the data points in the scatter plot was lowest for DNN and highest for MLR model.

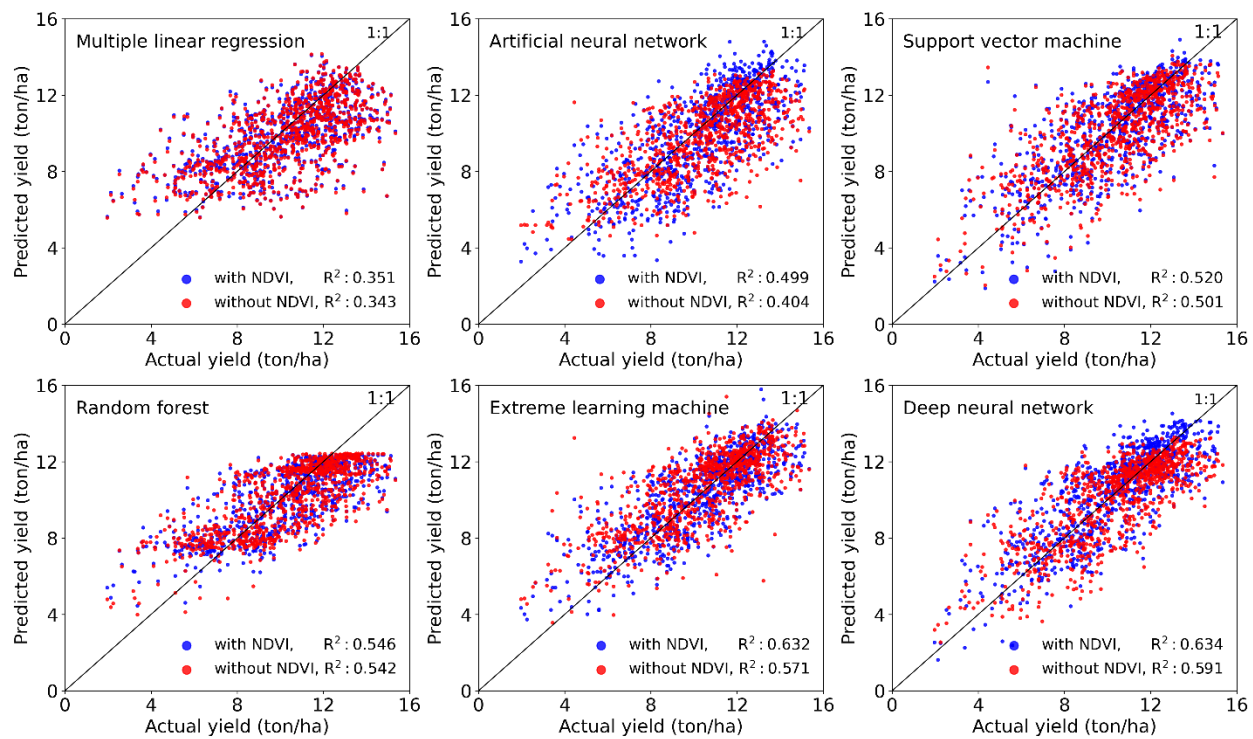


Figure 3-4. Crop yield estimates by the six machine learning models versus the actual crop yield obtained from the USDA NASS. Each subplot is for model using both input combinations.

Our findings indicate all adopted machine learning models have potential for predicting corn yield depending on the type of weather and satellite data used as input parameters. Among all the implemented models, results demonstrated that DNN showed highest predictive capabilities for estimating corn yield. Using weather and satellite data as an input feature increases the predictive power of the DNN model as compared to using only weather data.

Similar results were obtained in a study conducted by Kim & Lee, 2016. They found that DNN showed highest performance for predicting corn yield when compared to the SVM, RF and extremely randomized trees using climate and satellite data from May-September for the whole growing season in Iowa. The current study used additional satellite and weather data and found better performance using different machine learning models. Similarly, DNN when compared with MARS, SVM, RF, ERT, and ANN showed better performance for predicting corn yield using weather and satellite data in the midwestern United States (Kim et al., 2020). In addition, DNN outperformed LASSO, shallow neural network, and RT model when used for estimating corn yield using crop genotype, weather, and soil data across the United States and Canada (Khaki et al., 2020). Results confirms that DNN performs better than SVM, RF, and ANN for predicting corn yield over the Corn Belt in the United States.

Results demonstrated that ELM slightly showed weaker performance compared to DNN and outperformed all the other models. Previous studies (e.g. Kouadio et al., 2018; Maimaitiyiming et al., 2019; Taherei Ghazvinei et al., 2018) used ELM to predict coffee yield, grapevine berry yield and sugarcane growth. They found that ELM performance was better than MLR, RF, ANN, and genetic programming. However, the application of ELM for predicting corn yield is very limited previously and this study provides new results for applying ELM. Results shows that RF indicated better performance after DNN, and ELM and outperformed MLR, ANN, and SVM. RF showed better performance than SVM and lower performance than DNN when used for corn yield prediction in Iowa (Kim & Lee, 2016). RF model was compared with other machine learning and ensemble models to predict corn yield in the United States Corn Belt (Shahhosseini et al., 2021). They found that weather information alone is not sufficient, and models need more hydrological inputs to make improved yield predictions. RF was good

predictive model as compared to the SVM when used for corn yield prediction in the midwestern United States (Kim et al., 2020). RF model outperformed MLR when used to predict wheat, maize, potato at global and regional in the United States (Jeong et al., 2016).

SVM showed better performance after DNN, ELM, and RF and outperformed MLR and ANN. SVM showed 97% accuracy outperforming ANN with 93% accuracy when used for predicting wheat, rice, soybean, maize, millets, and pea yield using temperature, rainfall, location, soil type, pH, area, and relative humidity data (Agarwal & Tarar, 2021). Cai et al., 2019 also found that SVM outperformed ANN for predicting wheat yield using satellite and climate data in Australia. SVM was least accurate in predicting corn yield as compared to deep learning using different input combinations from remotely sensed enhanced vegetation index (EVI) and climate data (temperature, potential evapotranspiration, pressure) in Illinois (Kuwata & Shibasaki, 2015). SVM presented the most accurate results with the lowest average RMSE as compared to ANN and other machine learning models when used for predicting paddy yield in South Korea, and corn and soybean yield using satellite and weather data in Illinois and Iowa, United States (Ju et al., 2021). SVM was compared with ANN and genetic programming to predict crop yield under climate change conditions in eastern Iran (Arefinia et al., 2021). They found that SVM model achieved highest accuracy among all the models.

Our study showed that ANN performance was lower as compared to all other models except MLR. Similar results were obtained by Kaul et al., 2005. They compared ANN and MLR models to predict corn and soybean yield using different input variables from rainfall and soil data in Maryland. They found that ANN performance was better than MLR. Gandhi et al., 2016 applied ANN model to predict rice yield using precipitation, temperature, and evapotranspiration data. They found that ANN can predict rice yield with 97.54% accuracy in different districts of

Maharashtra, India. Non-linear model bayesian neural network showed slightly better performance than MLR when used for predicting crop yield (barley, canola, spring wheat) using NDVI and EVI data in Canada (Johnson et al., 2016). Gopal & Bhargavi, 2019 applied ANN, MLR, SVM, KNN and hybrid MLR-ANN models to predict paddy yield using agricultural production and weather data in Tamilnadu, India. They found that hybrid MLR-ANN model was better than all other models. However, ANN was slightly better than MLR. Abrougui et al., 2019 compared MLR and ANN predictive abilities for estimating organic potato yield using soil properties and tillage data in Tunisia. They found that generally, the ANN model showed greater potential for predicting potato yield. ANN and MLR were implemented for forecasting maize yield using precipitation and fertilizer data in Jilin province of China (Matsumura et al., 2015). They found that ANN models significantly outperformed the MLR. ANN showed better performance as compared to MLR when used to predict corn and soybean yield on a county-by-county basis, in the corn belt area in the midwestern and Great Plains regions of the United States using NDVI data (A. Li et al., 2007).

Summary and conclusion

Accurate prediction of corn yield is crucial for farmers and decision makers for better planning and management. The United States supplies approximately 40% of global annual corn production and hence claimed to be the largest grain producer and exporter (Taylor & Koo, 2011). Developing reliable models to predict corn yield helps to improve food security and economy. In this study, application of six machine learning algorithms, including SVM, ANN, ELM, RF, MLR, and DNN were evaluated for predicting corn yield using meteorological and satellite-based data from 2003-2020 during the growing season (May through September). Analyses were conducted across the states in the midwestern United States, Kansas with 105

counties, Nebraska with 93 counties, and Iowa with 99 counties, which are part of the Corn Belt. All models were trained using “with NDVI” and “without NDVI” input combination. The “without NDVI” input combination includes only weather data of day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation. The “with NDVI” input combination includes all the parameters of “without NDVI” input plus NDVI information. Results indicated that performance of all models were slightly higher using “with NDVI” input combination as compared to using “without NDVI” input combination. There is approximately up to 23% increase in predictive capabilities of all applied models when NDVI information was used. This indicate that using NDVI increases the predictive power of the machine learning model for estimating corn yield. DNN showed the best performance following by ELM. MLR showed the worst performance. ELM was applied for predicting corn yield. Findings of this study could outbreak the limitation of ELM model to predict corn yield across the Corn Belt in the United States. Performance of the applied models can be further improved using soil related data in addition to climate and remote sensing data (Shahhosseini et al., 2021; Sharifi, 2021) for predicting crop yield. Machine learning can be built to predict crop yield at any time during the growing period.

Chapter 4 - Summary and Conclusion

The goal of this study was to answer the question: can machine learning models be able to predict ET_0 and corn yield across the central United States and the Corn Belt with reasonable accuracies? To address this question, different machine learning models were implemented to estimate ET_0 and crop yield. The model performance was evaluated using different statistical metrics of R^2 , RMSE, MAE, and RAE.

For predicting ET_0 , four different machine learning models (SVM, RF, ELM, and ANN) were implemented using data from 1979-2014 during the warm season (May through September) from 39 weather stations across the central United States. Four different input combination from meteorological data of minimum and maximum temperature (T_{min} and T_{max}), net solar radiation (R_n), relative humidity (H_r), and wind speed (U_2) were used as a input parameters to the models. Two strategies were used to develop the models and measure their efficiency: (i) General case: models were developed using data from all the weather stations (ii) Cluster case: weather stations were clustered/grouped with similar climatic characteristics, using K-means algorithms. Then models were developed for each group using data from weather stations in that group. All models were trained and tested using 1979-2004 and 2005-2014 datasets, respectively. Results in the first strategy revealed that all implemented models showed good performance with $R^2 > 0.961$ when five parameters of T_{max} , T_{min} , R_n , U_2 , and H_r were used for input variables. ELM performed best with an $R^2 = 0.976$ and $RMSE = 0.363 \text{ mmd}^{-1}$. When limited meteorological parameters, including only T_{max} , R_n , and H_r were used, all models still showed satisfactory results with an R^2 not less than 0.903. ELM performed best with an $R^2 = 0.917$ and $RMSE = 0.678 \text{ mmd}^{-1}$. Second strategy also revealed similar results with ELM showing the best performance using all four input combinations with RF indicating the worst performance. In addition, ELM

showed the smallest computational costs. Considering prediction accuracy, reliability, and computational cost of the studied models, ELM can be recommended when used with all five parameters or excluding a few parameters across the central United States.

For predicting corn yield, six different machine learning models of SVM, RF, ELM, ANN, MLR, DNN were developed using data from 2003-2020 during the growing season (May through September). Analyses were conducted across three states of Kansas, Nebraska, and Iowa in the Corn Belt, United States. All models were trained using “with NDVI” and “without NDVI” input combinations. The “without NDVI” input combination includes only weather data of day length, snow water equivalent, vapor pressure, solar radiation, maximum and minimum temperature, and precipitation. The “with NDVI” input combination includes all parameters of “without NDVI” input combination plus satellite based NDVI. Results revealed that all implemented models showed R^2 ranging from 0.351-0.634 when used with “with NDVI” input combination and R^2 ranging from 0.343-0.591 when used with “without NDVI” input combination. This showed that performance of all models were slightly higher using the “with NDVI” input combination as compared to using the “without NDVI” input combination. There is less than 19% decrease in the predictive capabilities of the applied models not using NDVI information. This indicate that using only weather data can provide satisfactory performance. In addition, DNN showed the best performance following by ELM. MLR showed the worst performance using both input combinations. This result could outbreak the limitation of ELM model to predict corn yield across the Corn Belt in the United States.

The first study shows the accuracy lies between R^2 values of 0.977 to 0.703 and second study shows the accuracy lies between R^2 values of 0.343 to 0.634. The above two studies indicate that machine learning algorithms could be very useful tool for estimating ET_0 and crop

yield. However, the performance of models largely depends upon on the types of algorithm and input parameters used. The results of this study could help choosing the best machine learning model when predicting ET_0 which helps in conservation of water resource and irrigation water management. Information on crop yields before harvest helps farmers and decision makers for better planning and management which increases economy and enhance food security.

Machine learning model might not work for all the climatic conditions. That is developing machine learning models for one climatic condition and using it for another climatic condition might not perform best. Therefore, a future step to this work could be to test the reliability of the machine learning models in different climatic conditions and then improve their performance using long period data from different climatic conditions. Increase the performance of the models using limited agriculture data availability.

Chapter 5 - References

- Abdipour, M., Younessi-Hmazekhanlu, M., & Ramazani, S. H. R. (2019). Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (*Carthamus tinctorius* L.). *Industrial Crops and Products*, 127, 185–194.
- Abdullah, S. S., Malek, M. A., Abdullah, N. S., Kisi, O., & Yap, K. S. (2015). Extreme learning machines: a new approach for prediction of reference evapotranspiration. *Journal of Hydrology*, 527, 184–195.
- Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., & Wilby, R. L. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*. <https://doi.org/10.1177/0309133312444943>
- Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., & Chehaibi, S. (2019). Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). *Soil and Tillage Research*, 190, 202–208.
- Adisa, O. M., Botai, J. O., Adeola, A. M., Hassen, A., Botai, C. M., Darkey, D., & Tesfamariam, E. (2019). Application of artificial neural network for predicting maize production in South Africa. *Sustainability*, 11(4), 1145.
- Adnan, M., Ahsan, M., -, A.-R., & Nazir, M. (2017). Estimating Evapotranspiration using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, 8(9), 108–113. <https://doi.org/10.14569/ijacsa.2017.080915>
- Agarwal, S., & Tarar, S. (2021). a Hybrid Approach for Crop Yield Prediction Using Machine Learning and Deep Learning Algorithms. *Journal of Physics: Conference Series*, 1714(1), 12012.
- Allen, R G, Pereira, L. S., Raes, D., Smith, M., & Ab, W. (1998). Allen_FAO1998. *Food and Agriculture Organization: Rome, Italy*, 300.
- Allen, Richard G, Pereira, L. S., Raes, D., Smith, M., & Ab, W. (1998). Fao,1998. *Irrigation and Drainage Paper No. 56, FAO*, 300. <https://doi.org/10.1016/j.eja.2010.12.001>

- Alonso, J., Castañón, Á. R., & Bahamonde, A. (2013). Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and Electronics in Agriculture*, 91, 116–120.
- Althoff, D., Bazame, H. C., Filgueiras, R., & Dias, S. H. B. (2018). Heuristic methods applied in reference evapotranspiration modeling. *Ciencia e Agrotecnologia*, 42(3), 314–324.
<https://doi.org/10.1590/1413-70542018423006818>
- Anapalli, S. S., Ahuja, L. R., Gowda, P. H., Ma, L., Marek, G., Evett, S. R., & Howell, T. A. (2016). Simulation of crop evapotranspiration and crop coefficients with data in weighing lysimeters. *Agricultural Water Management*. <https://doi.org/10.1016/j.agwat.2016.08.009>
- Arefinia, A., Bozorg-Haddad, O., Ahmadaali, K., Bazrafshan, J., Zolghadr-Asli, B., & Chu, X. (2021). Estimation of geographical variations in virtual water content and crop yield under climate change: comparison of three data mining approaches. *Environment, Development and Sustainability*, 1–19.
- Bachour, R., Maslova, I., Ticlavilca, A. M., Walker, W. R., & McKee, M. (2016). Wavelet-multivariate relevance vector machine hybrid model for forecasting daily evapotranspiration. *Stochastic Environmental Research and Risk Assessment*, 30(1), 103–117.
- Bellido-Jiménez, J. A., Estévez, J., & García-Marín, A. P. (2021). New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. *Agricultural Water Management*, 245(June 2020). <https://doi.org/10.1016/j.agwat.2020.106558>
- Binch, A., & Fox, C. W. (2017). Controlled comparison of machine vision algorithms for Rumex and Urtica detection in grassland. *Computers and Electronics in Agriculture*, 140, 123–138.
- Bondre, D. A., & Mahagaonkar, S. (2019). Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 4(5), 371–376.
- Bose, P., Kasabov, N. K., Bruzzone, L., & Hartono, R. N. (2016). Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11), 6563–6573.

- Brdar, S., Culibrk, D., Marinkovic, B., Crnobarac, J., & Crnojevic, V. (2011). Support vector machines with features contribution analysis for agricultural yield prediction. *Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (EcoSense 2011), Belgrade, Serbia*, 43–47.
- Breiman, L. (2001). Random forests. *Machine Learning*.
<https://doi.org/10.1023/A:1010933404324>
- Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., & You, L. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, 274, 144–159.
- Cf, O. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. *United Nations: New York, NY, USA*.
- Chen, D. (2011). Daily reference evapotranspiration estimation based on least squares support vector machines. *International Conference on Computer and Computing Technologies in Agriculture*, 54–63.
- Chen, Z., Zhu, Z., Jiang, H., & Sun, S. (2020). Estimating daily reference evapotranspiration based on limited meteorological data using deep learning and classical machine learning methods. *Journal of Hydrology*, 591(April), 125286.
<https://doi.org/10.1016/j.jhydrol.2020.125286>
- Chia, M. Y., Huang, Y. F., & Koo, C. H. (2020). Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. *Computers and Electronics in Agriculture*, 175(June). <https://doi.org/10.1016/j.compag.2020.105577>
- Chia, M. Y., Huang, Y. F., & Koo, C. H. (2021). Swarm-based optimization as stochastic training strategy for estimation of reference evapotranspiration using extreme learning machine. *Agricultural Water Management*, 243(May 2020), 106447.
<https://doi.org/10.1016/j.agwat.2020.106447>
- Chia, M. Y., Huang, Y. F., Koo, C. H., & Fung, K. F. (2020). Recent Advances in Evapotranspiration Estimation Using Artificial Intelligence Approaches with a Focus on Hybridization Techniques—A Review. *Agronomy*, 10(1), 101.
<https://doi.org/10.3390/agronomy10010101>

- Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69.
- Córdova, M., Carrillo-Rojas, G., Crespo, P., Wilcox, B., & Céleri, R. (2015). Evaluation of the Penman-Monteith (FAO 56 PM) Method for Calculating Reference Evapotranspiration Using Limited Data. *Mountain Research and Development*, 35(3), 230–239.
<https://doi.org/10.1659/MRD-JOURNAL-D-14-0024.1>
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*.
<https://doi.org/10.1023/A:1022627411411>
- Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792.
- Dahikar, S. S., & Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 2(1), 683–686.
- Dang, C., Liu, Y., Yue, H., Qian, J., & Zhu, R. (2021). Autumn crop yield prediction using data-driven approaches:-support vector machines, random Forest, and deep neural network methods. *Canadian Journal of Remote Sensing*, 47(2), 162–181.
- Deines, J. M., Kendall, A. D., Butler, J. J., Basso, B., & Hyndman, D. W. (2021). Combining Remote Sensing and Crop Models to Assess the Sustainability of Stakeholder-Driven Groundwater Management in the US High Plains Aquifer. *Water Resources Research*.
<https://doi.org/10.1029/2020wr027756>
- Deo, R. C., Samui, P., & Kim, D. (2016). Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stochastic Environmental Research and Risk Assessment*, 30(6), 1769–1784.
<https://doi.org/10.1007/s00477-015-1153-y>
- Dodds, F., & Bartram, J. (2016). *The water, food, energy and climate Nexus: Challenges and an agenda for action*. Routledge.
- Doorenbos, J., & Kassam, A. H. (1979). Yield response to water. FAO. *Irrigation and Drainage Paper*, 33, 193–201.

- Dou, X., & Yang, Y. (2018). Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. *Computers and Electronics in Agriculture*, 148(March), 95–106. <https://doi.org/10.1016/j.compag.2018.03.010>
- Dunn, J. H. R., Willett, M. K., Parker, E. D., & Mitchell, L. (2016). Expanding HadISD: Quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5(2), 473–491. <https://doi.org/10.5194/gi-5-473-2016>
- Elavarasan, D., & Vincent, P. M. D. R. (2021). A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing*, 1–14.
- Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, 36(2), 27.
- Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., & Xiang, Y. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*, 263(September), 225–241. <https://doi.org/10.1016/j.agrformet.2018.08.019>
- Feng, Y., Cui, N., Gong, D., Zhang, Q., & Zhao, L. (2017). Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. *Agricultural Water Management*, 193, 163–173. <https://doi.org/10.1016/j.agwat.2017.08.003>
- Feng, Y., Peng, Y., Cui, N., Gong, D., & Zhang, K. (2017). Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Computers and Electronics in Agriculture*, 136, 71–78.
- Ferreira, L. B., da Cunha, F. F., de Oliveira, R. A., & Fernandes Filho, E. I. (2019a). Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2019.03.028>
- Ferreira, L. B., da Cunha, F. F., de Oliveira, R. A., & Fernandes Filho, E. I. (2019b). Estimation

- of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. *Journal of Hydrology*, 572(February), 556–570.
<https://doi.org/10.1016/j.jhydrol.2019.03.028>
- Filgueiras, R., Almeida, T. S., Mantovani, E. C., Dias, S. H. B., Fernandes-Filho, E. I., da Cunha, F. F., & Venancio, L. P. (2020). Soil water content and actual evapotranspiration predictions using regression algorithms and remote sensing data. *Agricultural Water Management*, 241(February), 106346. <https://doi.org/10.1016/j.agwat.2020.106346>
- Gandhi, N., Petkar, O., & Armstrong, L. J. (2016). Rice crop yield prediction using artificial neural networks. *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 105–110.
- Genaidy, M. A. (2020). Estimating of Evapotranspiration Using Artificial Neural Network. *Misr Journal of Agricultural Engineering*, 37(1), 81–94.
<https://doi.org/10.21608/mjae.2020.94971>
- Gocic, M., Petković, D., Shamshirband, S., & Kamsin, A. (2016). Comparative analysis of reference evapotranspiration equations modelling by extreme learning machine. *Computers and Electronics in Agriculture*. <https://doi.org/10.1016/j.compag.2016.05.017>
- Gong, D., Hao, W., Gao, L., Feng, Y., & Cui, N. (2021). Extreme learning machine for reference crop evapotranspiration estimation: Model optimization and spatiotemporal assessment across different climates in China. *Computers and Electronics in Agriculture*, 187, 106294.
- Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. *The Scientific World Journal*, 2014.
- Gonzalez del Cerro, R. T., Subathra, M. S. P., Manoj Kumar, N., Verrastro, S., & Thomas George, S. (2020). Modelling the daily reference evapotranspiration in semi-arid region of South India: A case study comparing ANFIS and empirical models. *Information Processing in Agriculture*. <https://doi.org/10.1016/j.inpa.2020.02.003>
- Gopal, P. S. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture*, 165, 104968.
- Granata, F. (2019). Evapotranspiration evaluation models based on machine learning

- algorithms—A comparative study. *Agricultural Water Management*, 217(August 2018), 303–315. <https://doi.org/10.1016/j.agwat.2019.03.015>
- Grinblat, G. L., Uzal, L. C., Larese, M. G., & Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, 127, 418–424.
- Hara, P., Piekutowska, M., & Niedbała, G. (2021). Selection of Independent Variables for Crop Yield Prediction Using Artificial Neural Network Models with Remote Sensing Data. *Land*, 10(6), 609.
- Hassan, M. A., Khalil, A., Kaseb, S., & Kassem, M. A. (2017). Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Applied Energy*, 203, 897–916. <https://doi.org/10.1016/j.apenergy.2017.06.104>
- Heddami, S., Kisi, O., Sebbar, A., Houichi, L., & Djemili, L. (2020). New Formulation for Predicting Daily Reference Evapotranspiration (ET₀) in the Mediterranean Region of Algeria Country: Optimally Pruned Extreme Learning Machine (OPELM) Versus Online Sequential Extreme Learning Machine (OSELM). *Water Resources in Algeria-Part I: Assessment of Surface and Groundwater Resources*, 181–199.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507.
- Howell, T. A. (1990). *Relationships between crop production and transpiration, evapotranspiration, and irrigation*.
- Huang, Guang Bin, Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3), 489–501. <https://doi.org/10.1016/j.neucom.2005.12.126>
- Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70(1–3), 489–501.
- Huang, Gao, Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks*, 61, 32–48.
- Irmak, S., Kabenge, I., Skaggs, K. E., & Mutiibwa, D. (2012). Trend and magnitude of changes

- in climate variables and reference evapotranspiration over 116-yr period in the Platte River Basin, central Nebraska-USA. *Journal of Hydrology*, 420–421, 228–244.
<https://doi.org/10.1016/j.jhydrol.2011.12.006>
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., & Reddy, V. R. (2016). Random forests for global and regional crop yield predictions. *PLoS One*, 11(6), e0156571.
- Jiang, D., Yang, X., Clinton, N., & Wang, N. (2004). An artificial neural network model for estimating crop yields using remotely sensed information. *International Journal of Remote Sensing*, 25(9), 1723–1732.
- Johann, A. L., de Araújo, A. G., Delalibera, H. C., & Hirakawa, A. R. (2016). Soil moisture modeling based on stochastic behavior of forces on a no-till chisel opener. *Computers and Electronics in Agriculture*, 121, 420–428.
- Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, 218, 74–84.
- Ju, S., Lim, H., Ma, J. W., Kim, S., Lee, K., Zhao, S., & Heo, J. (2021). Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models. *Agricultural and Forest Meteorology*, 307, 108530.
- Kasun, L. L. C., Zhou, H., Huang, G.-B., & Vong, C. M. (2013). *Representational learning with ELMs for big data*.
- Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, 85(1), 1–18.
- Keshtegar, B., Kisi, O., Arab, H. G., & Zounemat-Kermani, M. (2018). Subset modeling basis ANFIS for prediction of the reference evapotranspiration. *Water Resources Management*, 32(3), 1101–1116.
- Keshtegar, B., Piri, J., & Kisi, O. (2016). A nonlinear mathematical modeling of daily pan evaporation based on conjugate gradient method. *Computers and Electronics in Agriculture*.
<https://doi.org/10.1016/j.compag.2016.05.018>

- Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 621.
- Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10, 1750.
- Kim, N., & Lee, Y.-W. (2016). Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, 34(4), 383–390.
- Kim, N., Na, S.-I., Park, C.-W., Huh, M., Oh, J., Ha, K.-J., Cho, J., & Lee, Y.-W. (2020). An artificial intelligence approach to prediction of corn yields under extreme weather conditions using satellite and meteorological data. *Applied Sciences*, 10(11), 3785.
- Kisi, O., & Alizamir, M. (2018). Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: Wavelet extreme learning machine vs wavelet neural networks. *Agricultural and Forest Meteorology*, 263(March), 41–48.
<https://doi.org/10.1016/j.agrformet.2018.08.007>
- Kisi, O., Shiri, J., Karimi, S., & Adnan, R. M. (2018). Three different adaptive neuro fuzzy computing techniques for forecasting long-period daily streamflows. In *Big data in engineering applications* (pp. 303–321). Springer.
- Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., & Mushtaq, S. (2018). Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Computers and Electronics in Agriculture*, 155, 324–338.
- Kousari, M. R., Hosseini, M. E., Ahani, H., & Hakimelahi, H. (2017). Introducing an operational method to forecast long-term regional drought based on the application of artificial intelligence capabilities. *Theoretical and Applied Climatology*.
<https://doi.org/10.1007/s00704-015-1624-6>
- Kross, A., Znoj, E., Callegari, D., Kaur, G., Sunohara, M., van Vliet, L., Rudy, H., Lapen, D., & McNairn, H. (2018). Evaluation of an artificial neural network approach for prediction of corn and soybean yield. *Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada*, 24–27.
- Kung, H.-Y., Kuo, T.-H., Chen, C.-H., & Tsai, P.-Y. (2016). Accuracy analysis mechanism for

- agriculture data using the ensemble neural network method. *Sustainability*, 8(8), 735.
- Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 858–861.
- Landeras, G., Ortiz-Barredo, A., & López, J. J. (2008). Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). *Agricultural Water Management*, 95(5), 553–565. <https://doi.org/10.1016/j.agwat.2007.12.011>
- Letey, J., Cardon, G. E., & Kan, I. (2007). Irrigation efficiency and uniformity. *Irrigation of Agricultural Crops*, 30, 117–132.
- Li, A., Liang, S., Wang, A., & Qin, J. (2007). Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing*, 73(10), 1149–1157.
- Li, R., Li, C., Xu, X., Wang, J., Yang, X., Huang, W., & Pan, Y. (2009). Winter wheat yield estimation based on support vector machine regression and multi-temporal remote sensing data. *Transactions of the Chinese Society of Agricultural Engineering*, 25(7), 114–117.
- Liang, B., Liu, H., Quine, T. A., Chen, X., Hallett, P. D., Cressey, E. L., Zhu, X., Cao, J., Yang, S., & Wu, L. (2021). Analysing and simulating spatial patterns of crop yield in Guizhou Province based on artificial neural networks. *Progress in Physical Geography: Earth and Environment*, 45(1), 33–52.
- Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling and Software*, 25(8), 891–909. <https://doi.org/10.1016/j.envsoft.2010.02.003>
- Maimaitiyiming, M., Sagan, V., Sidike, P., & Kwasniewski, M. T. (2019). Dual activation function-based Extreme Learning Machine (ELM) for estimating grapevine berry yield and quality. *Remote Sensing*, 11(7), 740.
- Maione, C., Batista, B. L., Campiglia, A. D., Barbosa Jr, F., & Barbosa, R. M. (2016). Classification of geographic origin of rice by data mining and inductively coupled plasma

- mass spectrometry. *Computers and Electronics in Agriculture*, 121, 101–107.
- Majhi, B., & Naidu, D. (2021). Differential evolution based radial basis function neural network model for reference evapotranspiration estimation. *SN Applied Sciences*, 3(1). <https://doi.org/10.1007/s42452-020-04069-z>
- Malik, A., & Kumar, A. (2015). Pan evaporation simulation based on daily meteorological data using soft computing techniques and multiple linear regression. *Water Resources Management*, 29(6), 1859–1872.
- Malik, A., Kumar, A., & Kisi, O. (2018). Daily pan evaporation estimation using heuristic methods with gamma test. *Journal of Irrigation and Drainage Engineering*, 144(9), 4018023.
- Maroufpoor, S., Bozorg-Haddad, O., & Maroufpoor, E. (2020). Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. *Journal of Hydrology*, 588(April), 125060. <https://doi.org/10.1016/j.jhydrol.2020.125060>
- Matsumura, K., Gaitan, C. F., Sugimoto, K., Cannon, A. J., & Hsieh, W. W. (2015). Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. *The Journal of Agricultural Science*, 153(3), 399–410.
- Mattar, M. A. (2018). Using gene expression programming in monthly reference evapotranspiration modeling: a case study in Egypt. *Agricultural Water Management*, 198, 28–38.
- Matthews, S. G., Miller, A. L., Plötz, T., & Kyriazakis, I. (2017). Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Scientific Reports*, 7(1), 1–12.
- McGuire, S. (2015). FAO, IFAD, and WFP. The State of Food Insecurity in the World 2015: Meeting the 2015 International Hunger Targets: Taking Stock of Uneven Progress. Rome: FAO, 2015. *Advances in Nutrition*, 6(5), 623–624. <https://doi.org/10.3945/an.115.009936>
- McGuire, V. L. (2017). Water-Level and Recoverable Water in Storage Changes, High Plains Aquifer, Predevelopment to 2015 and 2013–15. *U.S. Geological Survey Scientific Investigations Report 2017–5040*.

- Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture*. <https://doi.org/10.1016/j.compag.2017.05.002>
- Moghaddamnia, a, Ghafari, M., Piri, J., & Han, D. (2008). *Evaporation Estimation Using Support Vector Machines Technique*. 14–22.
- Mohammadi, B., & Mehdizadeh, S. (2020). Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. *Agricultural Water Management*. <https://doi.org/10.1016/j.agwat.2020.106145>
- Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., & Gocic, M. (2015). Extreme learning machine based prediction of daily dew point temperature. *Computers and Electronics in Agriculture*, 117, 214–225.
- Moorhead, J. E., Marek, G. W., Gowda, P. H., Lin, X., Colaizzi, P. D., Evett, S. R., & Kutikoff, S. (2019). Evaluation of evapotranspiration from eddy covariance using large weighing lysimeters. *Agronomy*. <https://doi.org/10.3390/agronomy9020099>
- Moratiel, R., Bravo, R., Saa, A., Tarquis, A. M., & Almorox, J. (2019). Estimation of evapotranspiration by FAO Penman–Monteith Temperature and Hargreaves–Samani models under temporal and spatial criteria. A case study in Duero Basin (Spain). *Natural Hazards and Earth System Sciences Discussions*. <https://doi.org/10.5194/nhess-2019-250>
- Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotziou, G., Wiebenson, J., Bill, R., & Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, 152, 104–116.
- Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., & Ramon, H. (2004). Automatic detection of ‘yellow rust’ in wheat using reflectance measurements and neural networks. *Computers and Electronics in Agriculture*, 44(3), 173–188.
- Mutiibwa, D., & Irmak, S. (2013). AVHRR-NDVI-based crop coefficients for analyzing long-term trends in evapotranspiration in relation to changing climate in the U.S. High Plains. *Water Resources Research*, 49(1), 231–244. <https://doi.org/10.1029/2012WR012591>

- Nahvi, B., Habibi, J., Mohammadi, K., Shamshirband, S., & Al Razgan, O. S. (2016). Using self-adaptive evolutionary algorithm to improve the performance of an extreme learning machine for estimating soil temperature. *Computers and Electronics in Agriculture*, 124, 150–160.
- NASS, U. (2017). *Land values 2017 summary*.
- Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Computers and Electronics in Agriculture*, 163, 104859.
- Oguntunde, P. G., Lischeid, G., & Dietrich, O. (2018). Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis. *International Journal of Biometeorology*, 62(3), 459–469.
- Parisouj, P., Mohebzadeh, H., & Lee, T. (2020). Employing Machine Learning Algorithms for Streamflow Prediction: A Case Study of Four River Basins with Different Climatic Zones in the United States. *Water Resources Management*. <https://doi.org/10.1007/s11269-020-02659-5>
- Paswan, R. P., & Begum, S. A. (2013). *Regression and neural networks models for prediction of crop production 1*.
- Patil, A. P., & Deka, P. C. (2016). An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Computers and Electronics in Agriculture*, 121, 385–392. <https://doi.org/10.1016/j.compag.2016.01.016>
- Paul, M., Vishwakarma, S. K., & Verma, A. (2015). Analysis of soil behaviour and prediction of crop yield using data mining approach. *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 766–771.
- Prasad, N. R., Patel, N. R., & Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*, 29(2), 195–206.
- Reis, M. M., da Silva, A. J., Zullo Junior, J., Tuffi Santos, L. D., Azevedo, A. M., & Lopes, É. M. G. (2019). Empirical and learning machine approaches to estimating reference evapotranspiration based on temperature data. *Computers and Electronics in Agriculture*, 165(April), 104937. <https://doi.org/10.1016/j.compag.2019.104937>

- Rosenblat, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. <https://doi.org/10.1037/h0042519>
- Sagan, V., Maimaitijiang, M., Bhadra, S., Maimaitiyiming, M., Brown, D. R., Sidike, P., & Fritschi, F. B. (2021). Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174, 265–281.
- Sahu, S., Chawla, M., & Khare, N. (2017). An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 53–57.
- Seifi, A., & Riahi, H. (2020). Estimating daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ann, and gamma test-anfis models in an arid area of iran. *Journal of Water and Climate Change*.
<https://doi.org/10.2166/wcc.2018.003>
- Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, 11(1), 1–15.
- Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, 101(3), 891–896.
- Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat Crop Yield Prediction Using Deep LSTM Model. *ArXiv Preprint ArXiv:2011.01498*.
- Shiri, J., Nazemi, A. H., Sadraddini, A. A., Landaras, G., Kisi, O., Fakheri Fard, A., & Marti, P. (2014). Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. *Computers and Electronics in Agriculture*, 108, 230–241. <https://doi.org/10.1016/j.compag.2014.08.007>
- Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos One*, 16(6), e0252402.

- Shrestha, N. K., & Shukla, S. (2015). *Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment*. Agricultural and Forest Meteorology; Elsevier B.V.
<https://doi.org/10.1016/j.agrformet.2014.09.025>
- Sivanandhini, P., & Prakash, J. (2020). Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network. *International Journal of Innovative Science and Research Technology*, 5(5), 1092–1096.
- Smidt, S. J., Haacker, E. M. K., Kendall, A. D., Deines, J. M., Pei, L., Cotterman, K. A., Li, H., Liu, X., Basso, B., & Hyndman, D. W. (2016). Complex water management in modern agriculture: Trends in the water-energy-food nexus over the High Plains Aquifer. *Science of the Total Environment*, 566, 988–1001.
- Stone, M., & Rahmani, V. (2020). Analysis of long-term extreme wetness, hotness, and coldness patterns in Kansas, USA. *ASABE 2020 Annual International Meeting*.
<https://doi.org/10.13031/aim.202001239>
- Sudheer, K. P., Gosain, A. K., & Ramasastri, K. S. (2003). Estimating Actual Evapotranspiration from Limited Climatic Data Using Neural Computing Technique. *Journal of Irrigation and Drainage Engineering*. [https://doi.org/10.1061/\(asce\)0733-9437\(2003\)129:3\(214\)](https://doi.org/10.1061/(asce)0733-9437(2003)129:3(214))
- Sujatha, R., & Isakki, P. (2016). A study on crop yield forecasting using classification techniques. *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, 1–4.
- Suleiman, A. A., & Hoogenboom, G. (2007). Comparison of Priestley-Taylor and FAO-56 Penman-Monteith for Daily Reference Evapotranspiration Estimation in Georgia. *Journal of Irrigation and Drainage Engineering*, 133(2), 175–182. [https://doi.org/10.1061/\(asce\)0733-9437\(2007\)133:2\(175\)](https://doi.org/10.1061/(asce)0733-9437(2007)133:2(175))
- Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). County-level soybean yield prediction using deep CNN-LSTM model. *Sensors*, 19(20), 4363.
- Suresh, A., Manjunathan, N., Rajesh, P., & Thangadurai, E. (2020). Crop Yield Prediction Using Linear Support Vector Machine. *European Journal of Molecular & Clinical Medicine*, 7(6), 2189–2195.

- Tabari, H., Kisi, O., Ezani, A., & Hosseinzadeh Talaei, P. (2012). SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. *Journal of Hydrology*, 444–445, 78–89.
<https://doi.org/10.1016/j.jhydrol.2012.04.007>
- Taherei Ghazvinei, P., Hassanpour Darvishi, H., Mosavi, A., Yusof, K. bin W., Alizamir, M., Shamshirband, S., & Chau, K. (2018). Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. *Engineering Applications of Computational Fluid Mechanics*, 12(1), 738–749.
- Tavakol, A., & Rahmani, V. (2020a). Climate Change. In B. K. Paul (Ed.), *Natural Hazards and Disasters: From Avalanches and Climate Change to Water Spouts and Wildfires [2 volumes]* (pp. 22–33). ABC-CLIO.
- Tavakol, A., & Rahmani, V. (2020b). Temperature Extremes. In Bimal Kanti Paul (Ed.), *Natural Hazards and Disasters: From Avalanches and Climate Change to Water Spouts and Wildfires [2 volumes]* (pp. 199–210). ABC-CLIO.
- Tavakol, A., Rahmani, V., & Harrington, J. (2020a). Changes in the frequency of hot, humid days and nights in the Mississippi River Basin. *International Journal of Climatology*.
<https://doi.org/10.1002/joc.6484>
- Tavakol, A., Rahmani, V., & Harrington, J. (2020b). Probability of compound climate extremes in a changing climate: A copula-based study of hot, dry, and windy events in the central United States. *Environmental Research Letters*, 15(10). <https://doi.org/10.1088/1748-9326/abb1ef>
- Tavakol, A., Rahmani, V., & Harrington, J. (2020c). Temporal and spatial variations in the frequency of compound hot, dry, and windy events in the central United States. *Scientific Reports*, 10(1), 1–13. <https://doi.org/10.1038/s41598-020-72624-0>
- Taylor, R. D., & Koo, W. W. (2011). *2011 Outlook of the US and World Wheat Industries, 2010-2020*.
- Terliksiz, A. S., & Altýlar, D. T. (2019). Use of deep neural networks for crop yield prediction: A case study of soybean yield in lauderdale county, alabama, usa. *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 1–4.

- Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2014). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
- Timm, A. U., Roberti, D. R., Streck, N. A., de Gonçalves, L. G. G., Acevedo, O. C., Moraes, O. L. L., Moreira, V. S., Degrazia, G. A., Ferlan, M., & Toll, D. L. (2014). Energy partitioning and evapotranspiration over a rice paddy in Southern Brazil. *Journal of Hydrometeorology*. <https://doi.org/10.1175/JHM-D-13-0156.1>
- Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709.
- Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. In *Renewable Energy*. <https://doi.org/10.1016/j.renene.2016.12.095>
- Wada, Y., & Bierkens, M. F. P. (2014). Sustainability of global water use: past reconstruction and future projections. *Environmental Research Letters*, 9(10), 104003.
- Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 1–5.
- Wang, S., Lian, J., Peng, Y., Hu, B., & Chen, H. (2019). Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. *Agricultural Water Management*. <https://doi.org/10.1016/j.agwat.2019.03.027>
- Wen, X., Si, J., He, Z., Wu, J., Shao, H., & Yu, H. (2015). Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration With Limited Climatic Data in Extreme Arid Regions. *Water Resources Management*, 29(9), 3195–3209. <https://doi.org/10.1007/s11269-015-0990-2>
- Wu, L., Peng, Y., Fan, J., Wang, Y., & Huang, G. (2021). A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation. *Agricultural Water*

Management, 245, 106624.

- Wu, L., Zhou, H., Ma, X., Fan, J., & Zhang, F. (2019). Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2019.123960>
- Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A. X., & Nemani, R. R. (2006). Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11), 3452–3461. <https://doi.org/10.1109/TGRS.2006.876297>
- Yaseen, Z. M., Al-Juboori, A. M., Beyaztas, U., Al-Ansari, N., Chau, K.-W., Qi, C., Ali, M., Salih, S. Q., & Shahid, S. (2020). Prediction of evaporation in arid and semi-arid regions: A comparative study using different machine learning models. *Engineering Applications of Computational Fluid Mechanics*, 14(1), 70–89.
- Yassin, M. A., Alazba, A. A., & Mattar, M. A. (2016). Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. *Agricultural Water Management*. <https://doi.org/10.1016/j.agwat.2015.09.009>
- Yin, Z., Wen, X., Feng, Q., He, Z., Zou, S., & Yang, L. (2017). Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area. *Hydrology Research*. <https://doi.org/10.2166/nh.2016.205>
- Yuan, X., Chen, C., Lei, X., Yuan, Y., & Adnan, R. M. (2018). Monthly runoff forecasting based on LSTM–ALO model. *Stochastic Environmental Research and Risk Assessment*, 32(8), 2199–2212.
- Zaefizadeh, M., Jalili, A., Khayatnezhad, M., Gholamin, R., & Mokhtari, T. (2011). Comparison of multiple linear regressions (MLR) and artificial neural network (ANN) in predicting the yield using its components in the hulless barley. *Advances in Environmental Biology*, 109–114.
- Zhou, Q., & Zhao, Y. (2013). The design and implementation of intrusion detection system based on data mining technology. *Research Journal of Applied Sciences, Engineering and*

Technology, 5(14), 3824–3829. <https://doi.org/10.19026/rjaset.5.4531>

Zhu, B., Feng, Y., Gong, D., Jiang, S., Zhao, L., & Cui, N. (2020). Hybrid particle swarm optimization with extreme learning machine for daily reference evapotranspiration prediction from limited climatic data. *Computers and Electronics in Agriculture*, 173(October 2019), 105430. <https://doi.org/10.1016/j.compag.2020.105430>

Appendix A - Mean values of meteorological data for 39 weather stations with location

St No.	HadISD StationID	Latitude (N)	Longitude (E)	Elevation (m)	T _{max} (°C)	T _{min} (°C)	U ₂ (ms ⁻¹)	H _r (%)	R _n (wm ⁻² s ⁻¹)	ET ₀ (mmd ⁻¹)	Cluster
1	722530-12921	29.5	98.5	241	33	22	3.8	66	255	7.8	3
2	722540-13958	30.3	97.8	201	33	22	3.1	66	252	7.5	3
3	722550-12912	28.9	96.9	35	33	23	4.0	76	255	7.1	3
4	722560-13959	31.6	97.2	152	34	22	4.3	65	253	7.9	3
5	722590-03927	32.9	97.0	171	33	22	4.4	62	254	8.1	3
6	722630-23034	31.4	100.5	584	34	20	4.2	57	261	8.5	3
7	722650-23023	31.9	102.2	872	33	19	5.0	52	263	9.1	3
8	722660-13962	32.4	99.7	546	33	21	4.7	57	261	8.6	3
9	722700-23044	31.8	106.4	1,194	34	19	3.5	39	272	9.6	1
10	723510-13966	34.0	98.5	310	33	20	4.9	60	258	8.5	3
11	723530-13967	35.4	97.6	392	31	19	4.6	65	254	7.6	3
12	723560-13968	36.2	95.9	198	31	20	4.0	66	246	7.3	3
13	723630-23047	35.2	101.7	1,099	31	16	5.8	56	258	8.4	3
14	723650-23050	35.0	106.6	1,619	30	16	3.9	38	284	9.4	1
15	724500-03928	37.6	97.4	403	31	18	5.0	64	248	7.5	3
16	724510-13985	37.8	100.0	787	31	16	5.9	60	249	7.9	3
17	724560-13996	39.1	95.6	267	29	17	3.4	69	245	6.7	2
18	724580-13984	39.6	97.7	448	29	17	4.9	66	245	7.1	3
19	724585-93997	38.9	98.8	568	30	17	5.6	63	248	7.6	3
20	724650-23065	39.4	101.7	1,114	29	13	5.3	60	241	7.3	2
21	724660-93037	38.8	104.7	1,884	26	11	4.1	49	278	7.9	1
22	724760-23066	39.1	108.5	1,481	30	15	4.1	36	292	9.7	1
23	725520-14935	41.0	98.3	561	28	15	4.7	67	240	6.6	2
24	725560-14941	42.0	97.4	473	28	15	4.5	67	239	6.5	2
25	725620-24023	41.1	100.7	847	28	13	4.2	65	236	6.5	2
26	725640-24018	41.2	104.8	1,863	24	10	4.8	53	263	7.3	2
27	725660-24028	41.9	103.6	1,202	29	12	4.2	57	246	7.1	2

28	725760-24021	42.8	108.7	1,704	26	11	3.2	42	269	7.9	1
29	726510-14944	43.6	96.8	435	26	14	4.4	69	237	6.1	2
30	726540-14936	44.4	98.2	390	27	13	4.4	68	234	6.1	2
31	726620-24090	44.0	103.1	963	27	12	4.6	58	238	6.8	2
32	726770-24033	45.8	108.5	1,092	26	12	4.2	49	240	7.2	2
33	727530-14914	46.9	96.8	274	25	12	4.7	65	226	6.0	2
34	727640-24011	46.8	100.8	503	26	11	4.1	63	226	6.0	2
35	727680-94008	48.2	106.6	697	26	11	4.9	53	224	6.8	2
36	727720-24144	46.6	112.0	1,167	25	10	3.3	50	254	6.9	2
37	727730-24153	46.9	114.1	973	26	8	2.8	55	242	6.4	2
38	727750-24143	47.5	111.4	1,117	25	9	4.4	51	240	6.8	2
39	742300-24037	46.4	105.9	800	28	12	4.3	49	230	7.3	2