



Article

Predicting Green Water Footprint of Sugarcane Crop Using Multi-Source Data-Based and Hybrid Machine Learning Algorithms in White Nile State, Sudan

Rogaia H. Al-TaHER¹ , Mohamed E. Abuarab¹ , Abd Al-Rahman S. Ahmed², Mohammed Magdy Hamed³ , Ali Salem^{4,5,*} , Sara Awad Helalia¹, Elbashir A. Hammad⁶ and Ali Mokhtar^{1,7}

- ¹ Department of Agricultural Engineering, Faculty of Agriculture, Cairo University, Giza 12613, Egypt; rogaiahmohamed30@gmail.com (R.H.A.-T.); mohamed.aboarab@agr.cu.edu.eg (M.E.A.); sarah.awad.helalia@gmail.com (S.A.H.); ali.mokhtar@agr.cu.edu.eg (A.M.)
 - ² Department of Natural Resources, Faculty of African Postgraduate Studies, Cairo University, Giza 12613, Egypt; abdelrahman.ahmed@cu.edu.eg
 - ³ Construction and Building Engineering Department, College of Engineering and Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), B 2401 Smart Village, Giza 12577, Egypt; eng.mohammedhamed@aast.edu
 - ⁴ Civil Engineering Department, Faculty of Engineering, Minia University, Minia 61111, Egypt
 - ⁵ Structural Diagnostics and Analysis Research Group, Faculty of Engineering and Information Technology, University of Pécs, Boszorkány ut2, H-7624 Pecs, Hungary
 - ⁶ Department of Agricultural Engineering, Faculty of Agriculture, University of Khartoum, Khartoum 11115, Sudan; bashir.hammad58@gmail.com
 - ⁷ School of Geographic Sciences, Key Lab. of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
- * Correspondence: salem.ali@mik.pt.e.hu



Citation: Al-TaHER, R.H.; Abuarab, M.E.; Ahmed, A.A.-R.S.; Hamed, M.M.; Salem, A.; Helalia, S.A.; Hammad, E.A.; Mokhtar, A. Predicting Green Water Footprint of Sugarcane Crop Using Multi-Source Data-Based and Hybrid Machine Learning Algorithms in White Nile State, Sudan. *Water* **2024**, *16*, 3241. <https://doi.org/10.3390/w16223241>

Academic Editors: Winnie Gerbens-Leenes and S.D. Vaca Jimenez

Received: 26 September 2024
Revised: 23 October 2024
Accepted: 4 November 2024
Published: 11 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Water scarcity and climate change present substantial obstacles for Sudan, resulting in extensive migration. This study seeks to evaluate the effectiveness of machine learning models in forecasting the green water footprint (GWFP) of sugarcane in the context of climate change. By analyzing various input variables such as climatic conditions, agricultural data, and remote sensing metrics, the research investigates their effects on the sugarcane cultivation period from 2001 to 2020. A total of seven models, including random forest (RF), extreme gradient boosting (XGBoost), and support vector regressor (SVR), in addition to hybrid combinations like RF-XGB, RF-SVR, XGB-SVR, and RF-XGB-SVR, were applied across five scenarios (Sc) which includes different combinations of variables used in the study. The most significant mean bias error (MBE) was recorded in RF with Sc3 (remote sensing parameters), at $5.14 \text{ m}^3 \text{ ton}^{-1}$, followed closely by RF-SVR at $5.05 \text{ m}^3 \text{ ton}^{-1}$, while the minimum MBE was $0.03 \text{ m}^3 \text{ ton}^{-1}$ in RF-SVR with Sc1 (all parameters). SVR exhibited the highest R^2 values throughout all scenarios. Notably, the R^2 values for dual hybrid models surpassed those of triple hybrid models. The highest Nash–Sutcliffe efficiency (NSE) value of 0.98 was noted in Sc2 (climatic parameters) and XGB-SVR, whereas the lowest NSE of 0.09 was linked to SVR in Sc3. The root mean square error (RMSE) varied across different ML models and scenarios, with Sc3 displaying the weakest performance regarding remote sensing parameters (EVI, NDVI, SAVI, and NDWI). Effective precipitation exerted the most considerable influence on GWFP, contributing 81.67%, followed by relative humidity (RH) at 7.5% and Tmax at 5.24%. The study concludes that individual models were as proficient as, or occasionally surpassed, double and triple hybrid models in predicting GWFP for sugarcane. Moreover, remote sensing indices demonstrated minimal positive influence on GWFP prediction, with Sc3 producing the lowest statistical outcomes across all models. Consequently, the study advocates for the use of hybrid models to mitigate the error term in the prediction of sugarcane GWFP.

Keywords: sugarcane GWFP; climate parameters; remote sensing indices; machine learning models; single and hybrid models

1. Introduction

Sudan ranks among the three largest countries in Africa in terms of land area and holds significant importance globally due to its abundant water and fertile agricultural land, covering approximately one-third of its total area of 1,886,068 square kilometres [1]. Additionally, Sudan boasts a wealth of natural agricultural resources such as fertile lands and animal, mineral, plant, and water resources.

Sugarcane is a vital crop worldwide for sugar production and renewable bioenergy, typically cultivated in dry and semiarid regions. China, the world's fourth largest sugar producer following Brazil, India, and the European Union, collectively accounts for about 80% of global production, while the remaining 20% is derived from sugar beets mainly grown in temperate regions of the Northern Hemisphere, utilized as a raw material for bioethanol production for renewable energy [2]. In Sudan, sugarcane is a key cash crop used for sugar production and other goods for both local consumption and export [3].

The concept of water footprint refers to the amount of fresh water utilized in the production process, encompassing all stages of the production chain. This water footprint comprises three components: green, blue, and grey water footprints. The green water footprint pertains to the consumption of green water resources (rainwater that does not turn into runoff), while the blue water footprint denotes the consumption of blue water resources (surface and groundwater). Lastly, the grey water footprint concerns pollution and represents the volume of fresh water necessary to absorb pollutants based on natural background concentrations and prevailing water quality standards [4–6].

The quantity of water used throughout the crop growth period directly influences the production, serving as the ultimate objective of agricultural endeavours. The assessment of the relationship between water usage and crop yield through water footprint accounting proves to be a suitable method [7]. Water footprint (WF) investigations primarily aim to decrease the global average of freshwater consumption. It is projected that the water footprint may rise by as much as 22% due to climate variations and alterations in land use by the year 2090 [8].

Xu, Chen [9] emphasized that water footprint, scarcity, and productivity serve as the primary metrics for assessing sustainable irrigated agriculture. Nevertheless, it has been established that water footprint (WF) is more practical and reliable when compared to other methodologies. This WF comprises green and blue components, denoting precipitation and irrigation water, respectively. Through WF assessment, the sustainability of water resources can be gauged, and the correlation between water usage and crop yield can be examined [10]. The evaluation of WF necessitates data collection and the creation of inventory analyses to elucidate the link between water utilization and crop yield [11]. Green water evapotranspiration (ET_{green}) represents the rainwater stored in the soil available for crop evapotranspiration, akin to effective precipitation (P_{eff}) [12], while blue water evapotranspiration (ET_{blue}) pertains to irrigation water sourced from groundwater aquifers and surface water during the growing seasons [13].

Remote sensing is described as the discipline of acquiring information about objects or regions from a distance without direct physical contact. It serves as a tool for monitoring the earth's resources through space technology alongside ground observations [14]. Remote sensing (RS) technologies offer a diagnostic mechanism acting as an early warning system, enabling timely interventions by the agricultural sector to mitigate potential issues before they escalate and detrimentally affect crop productivity. Despite the availability of various RS options due to advancements in sensor technologies, data management, and analytics, the agricultural industry has yet to fully embrace RS technologies due to uncertainties regarding their adequacy, suitability, and techno-economic viability [15,16]. RS technologies can aid in making site-specific management decisions at different crop production stages, thereby optimizing crop yield while addressing concerns related to environmental impact, profitability, and sustainability [15,16].

Machine learning (ML) has become a specialized field within Artificial Intelligence (AI) that uses algorithms to extract insights from large datasets and apply this knowledge

for self-improvement in making accurate calculations or predictions [17]. The application of machine learning techniques is widely seen in predicting evaporation from water surfaces, evapotranspiration, and various factors related to water resources, hydrology, water quality, and reservoir operations [18,19].

Computational intelligence and machine learning methodologies have evolved for the assessment, quantification, monitoring, and prediction of crops. The reliability of machine learning approaches and computational tools has facilitated the generation of accurate, timely future forecasts. By analyzing and processing historical data, future predictions can be derived. This research underscores the assessment and utilization of machine learning for predicting crop yields [20,21].

ML has exhibited remarkable performance in diverse challenging tasks such as image categorization, facial recognition, parameter estimation, and natural language processing through learning intricate characteristics and connections from extensive training datasets. Recent investigations have delved into its application in yield prediction [22,23]. Machine learning models expedite swift and optimal decision-making. The ML framework entails training and evaluation to forecast result accuracy. Recent methodologies have highlighted the merits and demerits of approaches proposed in the last five years, also comparing different machine learning algorithms utilized in contemporary agriculture [24]. Everingham, Sexton [25] highlighted the effectiveness of the random forest (RF) model in predicting sugarcane yields early in the season.

The purpose of estimating or predicting the water footprint is to estimate the amount of water a plant needs to produce a unit weight of the crop and is calculated as ($\text{m}^3 \text{kg}^{-1}$). Accordingly, following sustainable water management systems for irrigation water reduces the water footprint of a single crop, which allows the use of the water that has been saved to increase the cultivated area and achieve food security. This is in general. As for countries that rely on rainwater for irrigation, estimating the green water footprint contributes significantly to determining the amount of water consumed for each crop. Accordingly, in the event of drought or lack of rainfall, water footprint calculations allow focusing on crops with a high economic return and a low water footprint while importing crops with a high water footprint from countries that do not suffer from drought at a low price, which achieves sustainable water and economic management for the country and integration between different countries.

Limited research has addressed the sustainable management of current water resources by integrating the concept of water footprint with the use of remote sensing indicators to monitor plant health, soil salinity, and plant water stress over a period of about 20 years using machine learning algorithms to analyze large datasets and provide recommendations to combat climate change and promote sustainable water management. In addition, there is a lack of research conducted on sugarcane cultivation areas in Sudan, and most of it focused more on irrigation methods and did not link the green water footprint with remote sensing indicators and machine learning algorithms to predict the green water footprint of sugarcane, which enables sustainable water management of irrigation water and its preservation to increase the agricultural area and achieve food security in Sudan.

Therefore, the primary objectives of this study are as follows: (i) assess the response of sugarcane's green water footprint (GWFP) to climate variations over the period from 2001 to 2020; (ii) utilize four remote sensing indices to monitor the current status of sugarcane throughout the same timeframe and their impact on GWFP; (iii) develop and compare three machine learning models (SVM, RF, and XGB) individually and in a hybrid form over sugarcane; (iv) identify the most effective model under optimal conditions, achieving high accuracy and minimal error in predicting the GWFP of sugarcane. This investigation can thus introduce an innovative modelling approach that will bolster endeavours to address GWFP forecasting, which helps in implementing strategies to mitigate issues like water usage regulations and enhancing food safety protocols.

2. Materials and Methods

2.1. Study Area and Workflow

The field experiment was carried out in White Nile State, Sudan, focusing on sugar cane, as this region stands out as one of the primary sugarcane cultivation areas in the country (Figure 1). Situated in the southern part of Sudan, the White Nile State spans latitudes 12.00 to 13.30 ° N and longitudes 31.00 to 33.30 ° E and is at an elevation of 384 m above sea level, covering an area of 39,701 square kilometres. Characterized by an arid to semi-arid climate, the state experiences varying annual rainfall levels ranging from 300 mm in the north to over 600 mm in the south. White Nile State was specifically chosen due to its status as a leading sugarcane producer in Sudan.

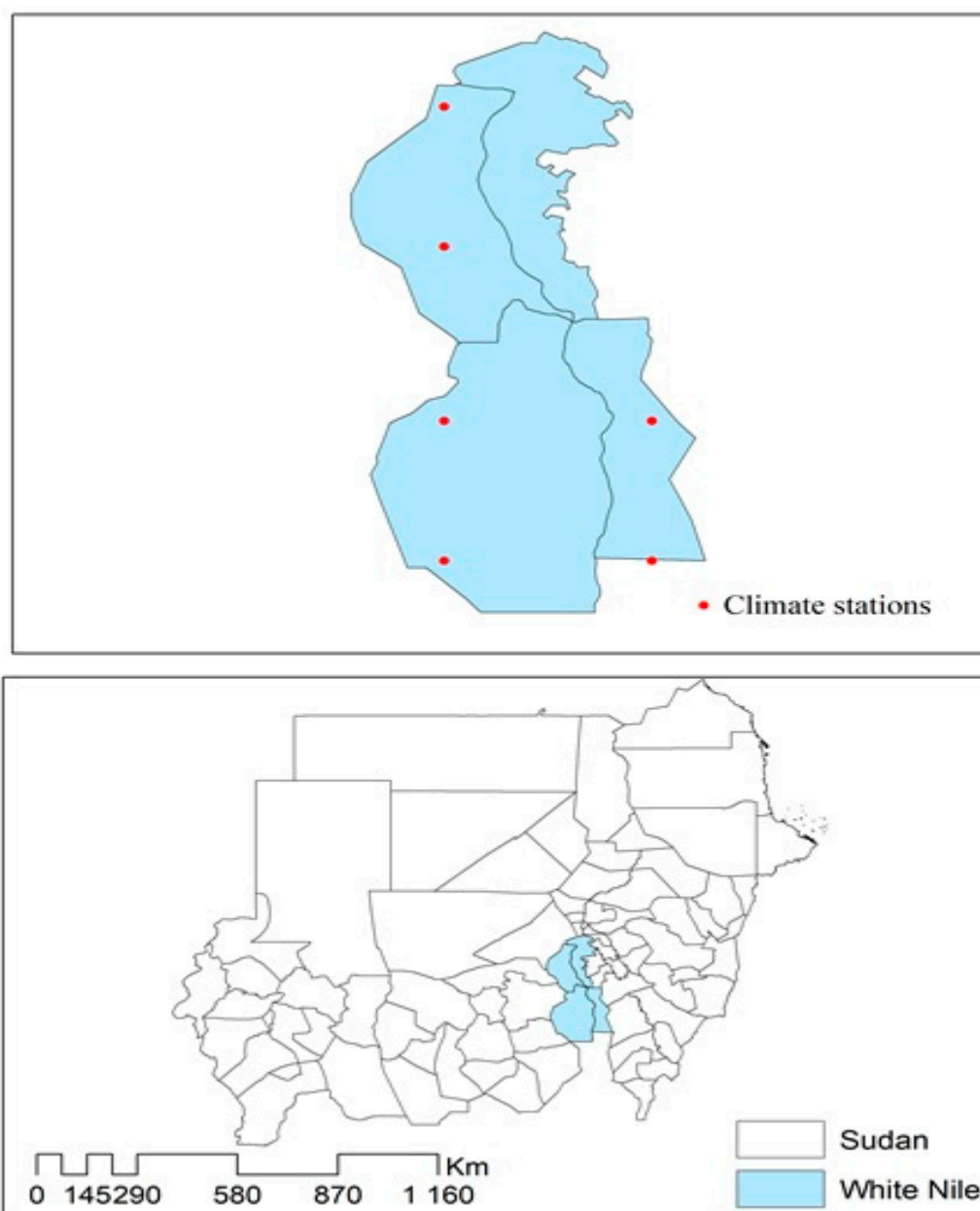


Figure 1. Geographical location of the study area and the meteorological stations.

The methodology of this research is depicted in Figure 2. The initial phase of the methodology involves the gathering of climate and crop data. Subsequently, three machine learning models (SVR, RF, and XGB) were utilized independently and in a hybrid form,

incorporating four remote sensing indices (EVI, NDVI, SAVI, and NDWI) to forecast the GWFP based on five scenarios that combine climate, crop data, and remote sensing indices.

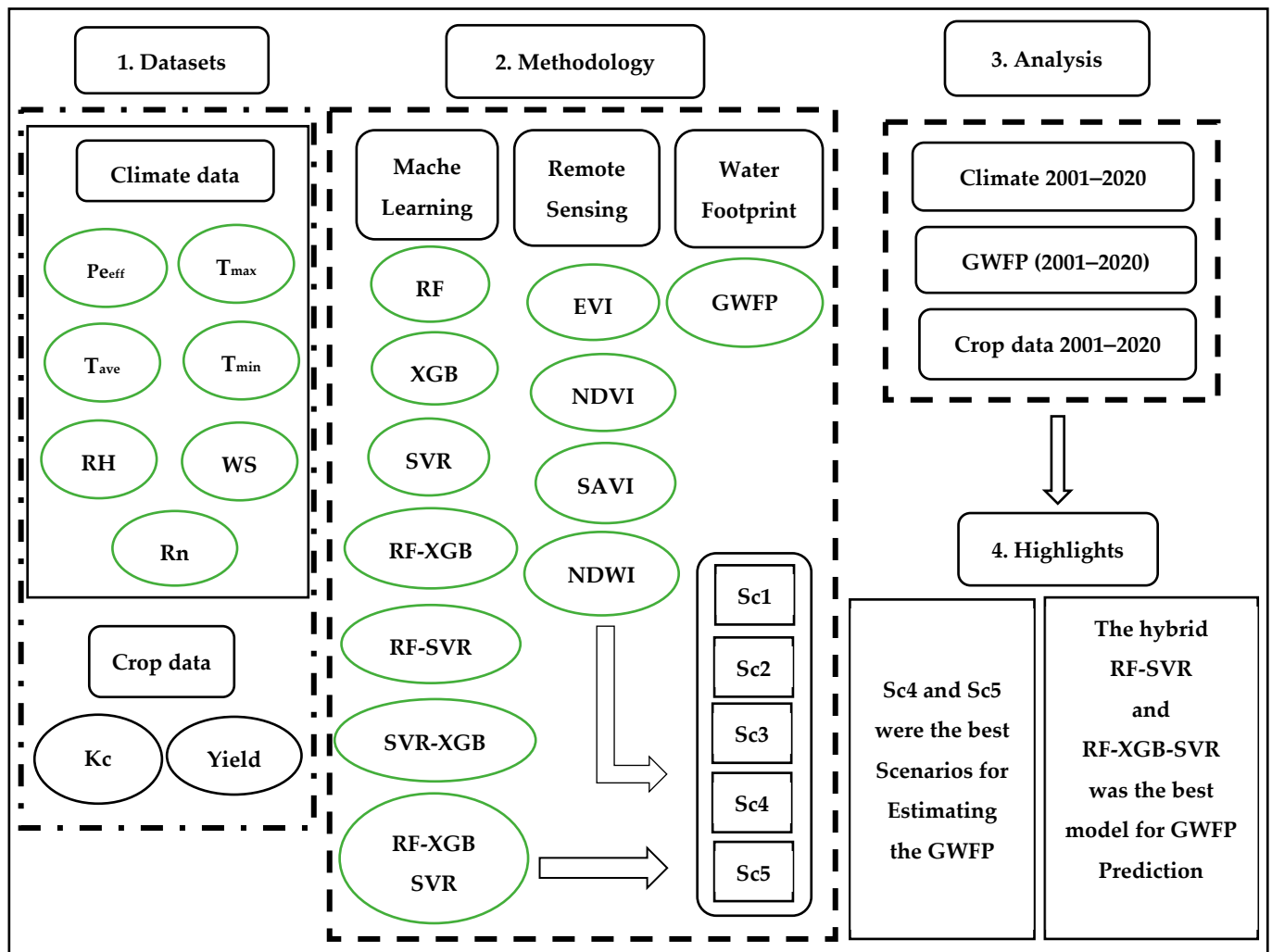


Figure 2. Workflow of the research. Note: T_{max} : maximum temperature; T_{min} : minimum temperature; RH: relative humidity; WS: wind speed; P_{eff} : effective precipitation; Kc: crop coefficient; SVR: support vector regression; RF: random forest; XGB: extreme gradient boosting; EVI: Enhanced Vegetation Index; NDVI: the normalized difference vegetation index; SAVI: soil-adjusted vegetation index; NDWI: the normalized difference water index; Sc: scenario.

2.2. Climate Conditions

The climate data encompassed monthly readings of minimum and maximum air temperature (T_{max} and T_{min} in $^{\circ}\text{C}$), wind speed (WS in ms^{-1}), relative humidity (RH in %), and precipitation (P in mm) collected between 2001 and 2020 from freely accessible data sourced from the NASA website, with a high-resolution daily time series of 0.5×0.5 degrees [26,27]. Additionally, solar radiation (SR) and vapour pressure deficit (VPD) data were obtained from <https://climate.northwestknowledge.net/> accessed on 1 March 2024. [28].

2.3. Remote Sensing Calculations and Field Measurements

This study employed Landsat time series data and GEE cloud computing to estimate four vegetation indices, with the GEE cloud computing process proving to be both rapid and effective. The computation of four remote sensing vegetation indices (EVI, NDVI, SAVI, and NDWI) was conducted using Landsat 7 and 8 Images through the Google Earth

Engine (GEE) (<https://earthengine.google.com/>) at a spatial resolution of 30 m at data level 2, with temporal resolution limited to the period from 2001 to 2020, focusing on the same season from April to November (Table 1). The Google Earth Engine (GEE) functions as a collaborative platform developed by Google, Carnegie Mellon University, and the United States Geological Survey, offering access to a wide array of functions and a vast archive of global satellite imagery and thematic maps spanning nearly four decades [29].

Table 1. Vegetation indices and remote sensing data description.

Index	Platform	Spatial Resolution (m)	Temporal Resolution (d)	Data Level	Years
EVI	Landsat7 ETM + Sensor Landsat8 OLI Sensor	30	2	L2	2001–2005, 2006–2010, 2011–2015, 2016–2020
NDVI	Landsat7 ETM + Sensor Landsat8 OLI Sensor	30	2	L2	2001–2005, 2006–2010, 2011–2015, 2016–2020
SAVI	Landsat7 ETM + Sensor Landsat8 OLI Sensor	30	2	L2	2001–2005, 2006–2010, 2011–2015, 2016–2020
NDWI	Landsat7 ETM + Sensor Landsat8 OLI Sensor	30	2	L2	2001–2005, 2006–2010, 2011–2015, 2016–2020

2.3.1. Multi-Temporal Image Analysis

Soil-Adjusted Vegetation Index (SAVI)

The SAVI is a commonly recognized and robust method utilized for vegetation delineation by exploiting the unique absorption property in the red spectrum and the high reflectance in the near-infrared (NIR) spectrum [30]. These specific spectral regions correspond to bands 3 and 4 in the Enhanced Thematic Mapper Plus (ETM+) and bands 4 and 5 in the Operational Land Imager (OLI) sensor image datasets. An advantageous aspect of employing SAVI is its normalization to a standardized reference point, with values ranging from -1 to 1 , ensuring comparability among SAVIs derived from different images. SAVI was applied to ETM+ and OLI images in the White Nile study region to assess vegetation cover density as per Qi, Chehbouni [31].

$$SAVI = \frac{NIR - RED}{(NIR + RED + L)} \times (1 + L) \quad (1)$$

where L is the soil-brightness correction factor ranging from 0 to 1 . In this study, L was 0.5 by default.

Normalized Difference Vegetation Index (NDVI)

The NDVI is the most widely used vegetation index, originally introduced by Rouse [32]. It can be expressed mathematically as follows:

$$NDVI = \frac{NIR - IR}{NIR + IR} \quad (2)$$

Due to the normalization process in its computation, NDVI values range from -1 to 1 , displaying a heightened sensitivity to green vegetation, even in regions with limited vegetation cover.

The Normalized Difference Water Index (NDWI)

The NDWI quantifies variations in leaf water content by utilizing the near-infrared (NIR) and shortwave infrared (SWIR) spectral bands. Being sensitive to both plant water

content and water bodies, NDWI is frequently utilized for monitoring droughts, tracking yield reductions, assessing reservoir discharge, groundwater level decreases, etc. [33].

$$\text{NDWI} = \frac{(\text{NIR} - \text{SWIR})}{(\text{NIR} + \text{SWIR})} \quad (3)$$

Values above 0.5 are indicative of water bodies, while vegetation typically exhibits lower values, facilitating the differentiation between vegetation and water bodies.

The Enhanced Vegetation Index (EVI)

EVI is a valuable method in remote sensing for evaluating vegetation health and monitoring changes over time. Derived from satellite imagery, EVI offers a quantitative assessment of vegetation density and vigour. Unlike conventional vegetation indices like NDVI, EVI factors in aerosol scattering and canopy background reflectance, making it more suitable for heavily vegetated regions or areas affected by atmospheric disturbances. EVI values range from -1 to 1 , with higher values denoting healthier and denser vegetation. Through the analysis of EVI data, the formula for Enhanced Vegetation Index (EVI) analysis is provided by the following:

$$\text{EVI} = 2.5 \times \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED} + 1} \quad (4)$$

where NIR represents the near-infrared band reflectance, and Red represents the red band reflectance.

2.4. Green Water Footprint

The Penman–Monteith equation [33–36] was utilized to determine the reference evapotranspiration (ET_0), a method endorsed by the Food and Agriculture Organization (FAO) [37] and proven effective by [38,39]. This equation is widely preferred for assessing water footprints [40]. The ET_0 calculator software (<http://www.fao.org/land-water/databases-and-software/eto-calculator/en/>, version 3.2, September 2012, accessed on 15 April 2022) was employed along with meteorological data to compute the ET_0 . The computation of the green water footprint adhered to the guidelines outlined in FAO Irrigation and Drainage Paper No. 56 [41], involving the determination of daily crop evapotranspiration (ET_c , mm) and effective precipitation (Pe_{eff} , mm) throughout the growing season.

$$ET_c = K_c \times ET_0 \quad (5)$$

Reference evapotranspiration (ET_0) is calculated in mm d^{-1} , with the adjusted crop coefficient denoted as K_c . ET_0 is estimated using the Penman–Monteith equation through the ET_0 calculator software. Additional information on data and calculations for ET_0 can be found in Mokhtar, He [34]. Adjustments to K_c are made based on FAO guidelines when RH_{min} deviates from 45% or when U_2 exceeds or falls below 2.0 m s^{-1} .

$$K_{c \text{ adjusted}} = K_{c \text{ reference}} + [0.04(U_2 - 2) - 0.004(RH_{\text{min}} - 45)] \left(\frac{h}{3}\right)^{0.3} \quad (6)$$

Pe is the effective precipitation (mm) over the growing season, and it was calculated using the following equation:

$$Pe = \begin{cases} \frac{P(4.17-0.2P)}{4.17}, & P < 8.3 \\ 4.17 + 0.1P, & P \geq 8.3 \end{cases} \quad (7)$$

where P and Pe are the monthly precipitation and effective precipitation [20], respectively [42,43]. Effective precipitation (Pe) over the growing period was calculated using the following formula. The water footprint (WF) is categorized into green water footprint

classifications [40]. As per Hoekstra [4] terminology, the water footprint of the crop season (WF_c) is the summation of the green components (WF_{green}) and is typically expressed in m³ ton⁻¹, equivalent to L Kg⁻¹.

$$\text{The green Water Footprint (WF}_{\text{green}}) = \frac{\text{CWR}_{\text{green}}}{Y} = 10 \times \frac{\text{ET}_{\text{green}}}{Y} \quad (8)$$

$$\text{The green water footprint (WF}_{\text{green}}) = \text{Max}(0, \text{ETc} - \text{pe}) \quad (9)$$

$$\text{The green water footprint (WF}_{\text{green}}) = \text{Min}(0, \text{ETc} - \text{pe}) \quad (10)$$

WF_{green} denotes the green water footprint, Y represents crop yield (ton ha⁻¹), and CWR_{green} signifies green water utilization (m³ ha⁻¹); ET_{green} and ET_{green} denote green (effective precipitation) and green (evapotranspiration) water, respectively. ET_c stands for crop evapotranspiration during the growing season [44].

$$\text{CWR}_{\text{green}} = 10 \sum \text{ET}_{\text{green}} \quad (11)$$

2.5. Machine Learning Implementations

In order to quantify the green water footprint, this study will employ three machine learning models: support vector regression (SVR), random forest (RF), and extreme gradient boosting (XGBoost). Two important processes for the machine learning models are training and testing. The data will be divided into two groups: the first group (65%) will be used for “training” the model, and the second group (35%) will be used for “testing” the model, which will evaluate the accuracy of the results produced by the calibrated learning machines by comparing the expected family water footprint values with the actual calculated values.

2.5.1. Random Forest (RF)

An ensemble of decision trees with controlled variance serves as the foundation for the RF model, which was created by Breiman [45] and Chen, Zhu [46]. One kind of bootstrap assembly is a random forest regression. It works with random binary trees that employ bootstrapping, a technique that involves selecting a random subset of the training dataset from the raw dataset and using it to grow the model using a portion of the observations. According to Chutia, Borah [47] and Ghorbani, Deo [48], RF is made up of an ensemble of chosen independent decision trees (DTs) that are identically distributed. The calculating process and comprehensive data are available in [45,49]. The following parameters were used to train the RF: 100 for the batch size, 100 for the bag size percent, 0 for the maximum depth, 1 for the number of execution slots, 100 for the number of iterations, and 1 for the random seed.

2.5.2. Extreme Gradient Boosting (XGBoost)

The XGB algorithm introduced by Chen and Guestrin [50] presents a new approach to implementing the Gradient Boosting Machine through regression trees. It relies on the concept of “boosting”, which involves aggregating predictions from a group of “weak” learners to create a “strong” learner through iterative training strategies. The formula for predicting at step t is as follows:

$$f_i^{(t)} = \sum_{k=1}^t f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \quad (12)$$

where $f_t(x_i)$ is the learner at step t , $f_i(t)$ and $f_i(t-1)$ are the predictions at steps t and $t-1$, and x_i is the input variable. The XGB uses the following analytical equation to determine

the “goodness” of the model derived from the original function in order to prevent the overfitting issue without affecting the computing performance of the model:

$$\text{Obj}^{(t)} = \sum_{k=1}^n l(\bar{y}_i, y_i) + \sum_{k=1}^t \Omega(f_i) \quad (13)$$

where l is the loss function, n is the number of observations, and Ω is the regularization term, which is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (14)$$

where ω is the vector of scores in the leaves, λ is the regularization parameter, and γ is the minimum loss needed to further partition the leaf node. XGBoost is an advanced algorithm that enhances the gradient-boosting decision tree methodology, excelling in the efficient construction of boosted trees and supporting parallel computation. It categorizes these boosted trees into regression and classification trees. The algorithm’s core focus lies in optimizing the objective function, as highlighted in previous studies [51,52]. The primary objective of XGBoost is to improve prediction accuracy by leveraging insights from previous weak learners and introducing tailored weak learners to address and rectify residual errors. This iterative process of combining multiple learners leads to predictions that outperform those of individual learners.

2.5.3. Support Vector Regression (SVR)

SVR is a supervised learning algorithm that can also function as a regression model while retaining key characteristics such as maximal margin. SVR shares a similar theory with SVM in terms of classification methodology, with minor adjustments. The primary goal is error minimization by customizing the hyperplane to increase the tolerance limit, considering that a portion of the error is acceptable. The approximate function in the SVM algorithm is as follows:

$$f(x) = \sum_{i=1}^1 (\alpha_i + \alpha_i^*) kx_i, k + b \quad (15)$$

where $f(x)$ is the relationship between dependent and independent variables, (α_i, α_i^*) is the Lagrangian multipliers, (kx_i, k) is the kernel function, and $b =$ the function bias.

2.5.4. Hybrid Model Building

For the earlier generations, the following four hybrid combinations were used: RF-XGB, RF-SVR, XGB-SVR, and RF-XGB-SVR. By using hybrid models, the results were expected to be more accurate and the inaccuracy in calculating the sugarcane crop’s green water footprint was to be minimized. It was carried out with the earlier hybrid method.

A hybrid approach involving random forest (RF) and extreme gradient boosting (XGB), known as RF-XGB, has been implemented to enhance tree-based algorithms for predictive modelling. In the context of decision trees, weak learners are typically shallow trees, sometimes as small as decision stumps (trees with two leaves). Boosting continuously updates the weights of the training set based on previous weaker learners to improve the importance of misclassified data. The function within XGB aims to enhance the majority vote value in RF while also aiding in the formation of individual trees within the bagging process of RF. The RF-XGB hybrid improves the RF model by reducing errors in MBE, MSE, and MAE values, consequently increasing model accuracy. This improvement is crucial as RF often struggles with overfitting due to challenges in determining the optimal number of trees [53].

The suggested approach used a meta-heuristic methodology to train the membership function parameters and layer weights of the basic RF model in an effort to increase the model’s efficiency for GWFP simulation. The suggested approach combined an SVR with

an RF to create a hybrid RF-SVR model, in which the SVR was used to optimize the RF settings. It is necessary to perform the optimization in the new n-dimensional space (RF-SVR) before using the SVR for training of RF weight updates. Under this configuration, the RF is viewed as a whale in a d-dimensional space, where d is the sum of the weights and the bias of the RF, varying the local and global optima of the search. Weights are assigned at random throughout the weight optimization procedure. If the difference between the output of the support vector regression and the output related to the real inputs is less than a certain limit, the initial weights are then adjusted and modified in each repeat. Every hybrid model with two algorithm models underwent the same processes.

XGB, SVR, and RF are combined in the hybrid model. The hybrid model is divided into three sections. The first component selects and pre-processes the data, ranking the relevance of each variable using the random forest method. The RF model’s input data are then chosen according to the variable’s relevance. Because the input variables in the model had varying dimensions, normalizing them was important to improve computing efficiency, convergence accuracy, and estimation precision. Equation (16) was used to normalize the variables to do this [54].

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{16}$$

where X_{\min} and X_{\max} are the minimum and maximum of all input variable values; X is the measured values of all the input variables; X' is the normalized values of measured values. The model simulation is covered in the second section. It starts with the initialization of the XGB model’s hyper-parameters, such as the regulatory coefficient (C) and the RBF parameters μ and ϵ . The model then estimates GWFP and computes the relevant error after defining the optimization accuracy requirement.

The model moves on to the third section, which entails applying the SVR method to optimize the XGB model’s hyper-parameters if the intended goal is not accomplished. The goal of optimizing the SVR model and raising the precision of the GWFP estimate is accomplished by returning the SVR hyper-parameters with the highest fitness.

2.6. Input Combination and Performance Evaluation of the Models

To investigate the weights and relationships between the available data and GWFP, this study used five scenarios, each consisting of different combinations of crop data, temperature data, and remote sensing indices (Table 2).

Table 2. The summary of the scenarios applied in this study.

Scenario	Input Parameters												
	Pe	T _{max}	T _{min}	RH	T _{ave}	Rn	WS	Kc _{adj}	SA	EVI	NDVI	SAVI	NDWI
Sc1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sc2	✓	✓	✓	✓	✓	✓	✓						
Sc3										✓	✓	✓	✓
Sc4	✓	✓	✓						✓				
Sc5	✓	✓											

Notes: Pe: effective precipitation; T_{max}: maximum temperature; T_{min}: minimum temperature; RH: relative humidity; WS: wind speed; Kc_{adj}: adjusted crop coefficient; SA: sown area; EVI: Landsat Enhanced Vegetation Index; NDVI: the normalized difference vegetation index; SAVI: soil-adjusted vegetation index; NDWI: the normalized difference water index; Sc: scenario.

Two subsets of the data were created: one from 2001 to 2020 for training, and the other from. By contrasting the projected and actual GWFP values from the models with the testing data, the models’ performance was verified. The models that were used were evaluated using the Nash–Sutcliffe model efficiency (NSE), the root mean squared error (RMSE), the mean absolute error (MAE), and the mean bias error (MBE) [55]. Furthermore,

the mean average percentage error (MAPE), accuracy (A), and coefficient of determination (R^2) were employed.

The data from the twenty seasons were split into two subsets: 30% of the data were set aside for testing and 70% of the data were used for training. The applied models were assessed using the mean absolute error (MAE), the root mean square error (RMSE), and the mean bias error [56]. Additionally, the T-Statistic test (Tstat) and uncertainty with a 95% confidence level (U95) are used to assess significance.

$$R^2 = \left[\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\left(\sum_{i=1}^n (O_i - \bar{O})^2\right) \left(\sum_{i=1}^n (P_i - \bar{P})^2\right)}} \right]^2 \tag{17}$$

Higher R^2 values correspond to higher prediction accuracy, and lower RMSE values suggest superior model performance. The applicable models were evaluated using the mean bias error (MBE).

$$MBE = \frac{1}{n} \sum_{i=1}^n (O_i - P_i) \tag{18}$$

Mean absolute error (MAE) measures the average magnitude of errors in projections without considering their signs. The absolute deviations between calculated and predicted yields are averaged across the test sample [57,58].

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \tag{19}$$

Moreover, the accuracy (A) and the coefficient of determination (R^2) are as follows:

$$A = 1 - \text{abs}\left(\frac{P_i - O_i}{O_i}\right) \tag{20}$$

$$T \text{ sat} = \sqrt{\frac{(1 - n)MBE^2}{RMSE^2 - MBE^2}} \tag{21}$$

$$MSE = \frac{1}{n} \sum (P_i - O_i)^2 \tag{22}$$

$$CC = \frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\left(\sum_{i=1}^n (O_i - \bar{O})^2\right) \left(\sum_{i=1}^n (P_i - \bar{P})^2\right)}} \tag{23}$$

The efficiency coefficient (NSE) of the Nash–Sutcliffe model, a normalized statistic that measures the relative size of residual variance to data variance, was used in this study’s performance statistics. The accuracy of the models displayed in Table 3 is indicated by the range of the scatter index (SI) [59] and the Nash–Sutcliffe efficiency coefficient (NSE) value [55]. As per Downing, Greenberg [60], the mean average percentage error (MAPE) was established. Furthermore, the purpose of the 95% uncertainty interval for model deviations is to evaluate significant differences between the estimated and forecasted GWFP to provide additional insight into the model’s effectiveness, which is characterized as

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{P_i - O_i}{O_i} \right| * 100 \tag{24}$$

$$MARE = \frac{1}{N} \sum_{i=1}^n |RE|_i \tag{25}$$

$$U_{95} = 1.96 \sqrt{(SD^2 + RMSE^2)} \quad (26)$$

Table 3. The range of NSE and SI.

NSE	Classifications	SI	Classifications
NSE = 1	Perfect	SI < 0.1	Excellent
NSE > 0.75	very good	0.1 < SI < 0.2	Good
0.74 > NSE > 0.64	Good	0.2 < SI < 0.3	Fair
0.64 > NSE > 0.5	Satisfactory	SI > 0.3	Poor
NSE < 0.5	Unsatisfactory		

In this context, O_i and P_i represent the actual and forecasted values and denote the actual and projected mean values. The relative error (RE) and standard difference (SD) between estimated and calculated values are also considered.

3. Results and Discussion

3.1. The Spatiotemporal Changes in Climate Variables (2001–2020)

Weather conditions play a crucial role in managing water resources, with their fluctuations being observed over the period from 2001 to 2020 (Figure 3). The range of temperatures varied significantly during this time, with the lowest recorded for maximum temperature (T_{max}) in 2014 at 36.83 °C and the highest in 2015 at 38.40 °C, averaging at 37.53 °C. Similarly, the lowest and highest values for minimum temperature (T_{min}) were noted in 2018 (25.06 °C) and 2010 (26.0 °C), respectively, with an average of 25.49 °C (Figure 3A).

The maximum relative humidity (RH) valued ranged from 40.85% in 2015 to 50.52%, which was recorded in 2019, with an average value over the time period equal to 46.03%. While the trend of wind speed (WS) values has the maximum value in 2017 of 5.02 ms^{-1} , the lowest value of wind speed was recorded in 2007 with 4.39 m s^{-1} having a mean value of 4.66 m s^{-1} (Figure 3B).

The highest values of effective precipitation were observed in 2007 at 120.45 mm and in 2019 (75.04 mm), while the lowest value was recorded in 2015 at 23.94, with an average value equal to 48.0 mm, where the rainfall season extends from May to October, while the sugarcane growing season extends from May to November, and there was no rainfall in November recorded through the time series 2001–2020. In the same context, the lowest and highest E_{To} values were reported in 2007 with 5.01 mm d^{-1} and in 2017 with 6.16 mm d^{-1} , having an average of 5.60 mm d^{-1} (Figure 3C).

The variations in E_{Tc} , yield, and GWFP values related to sugarcane growing seasons over the time series from 2001 to 2020 were highly related to the climate changes which have a great impact on climatic parameters and so on crop evapotranspiration, yield, and subsequently, the GWFP, where the maximum E_{Tc} was achieved in 2017 at 64.72 $\text{m}^3 \text{ha}^{-1}$ followed by 60.11 $\text{m}^3 \text{ha}^{-1}$ in 2015, while the lowest value was recorded in 2007 at 47.30 $\text{m}^3 \text{ha}^{-1}$ with a mean value equal to 55.36 $\text{m}^3 \text{ha}^{-1}$. Consequently, the maximum yield was achieved in 2008 at 48.40 ton ha^{-1} $\text{m}^3 \text{ha}^{-1}$ followed by 48.20 ton ha^{-1} in 2004, while the lowest value was recorded in 2016 at 34.60 ton ha^{-1} with a mean value equal to 42.71 ton ha^{-1} . Accordingly, the GWFP had the maximum in 2007 of 25.46 $\text{m}^3 \text{ton}^{-1}$ $\text{m}^3 \text{ha}^{-1}$ followed by 19.19 $\text{m}^3 \text{ton}^{-1}$ in 2019, while the lowest value was recorded in 2017 at 5.90 $\text{m}^3 \text{ton}^{-1}$ with a mean value equal to 11.27 $\text{m}^3 \text{ton}^{-1}$ (Figure 4).

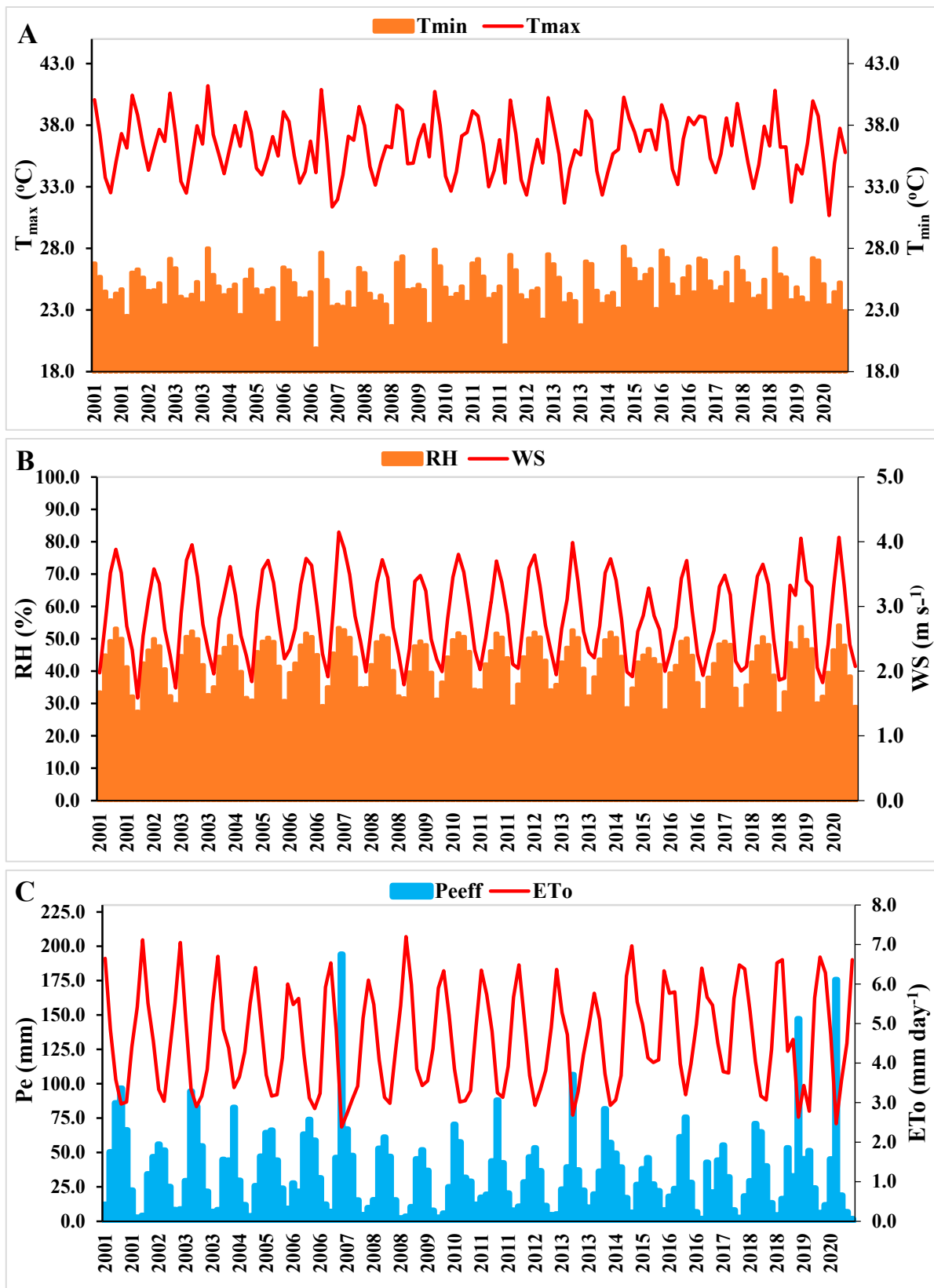


Figure 3. The climatic parameters and reference evapotranspiration from 2001 to 2020 in the study area are (A) Tmax and Tmin, (B) relative humidity and wind speed, and (C) effective precipitation and reference evapotranspiration.

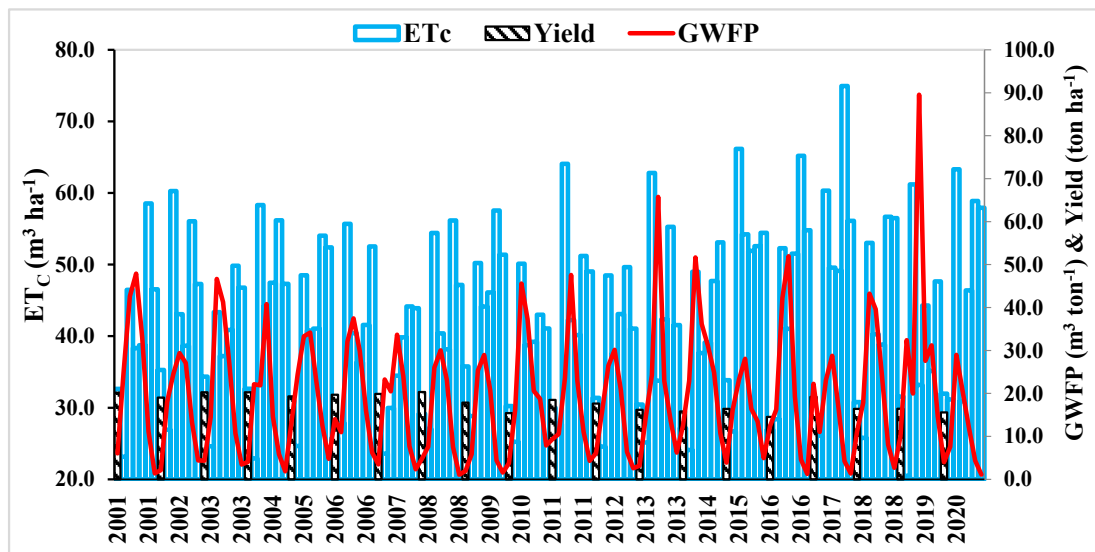


Figure 4. The evapotranspiration, yield, and green water footprint of sugarcane for the time series from 2001 to 2020.

3.2. The Spatiotemporal Changes in Vegetation Indices (2001–2020)

Figure 5 shows the locations of climate stations, each of which stands for a particular geographic area where climate data are gathered, as well as the average values of vegetation indices over 20 years. Four vegetation indices are linked to each station: the normalized difference water index (NDWI), the soil-adjusted vegetation index (SAVI), the Enhanced Vegetation Index (EVI), and the normalized difference vegetation index (NDVI).

Station	Index	Values
STAT-1	EVI	0.12
STAT-1	NDVI	0.17
STAT-1	SAVI	0.26
STAT-1	NDWI	-0.07
STAT-2	EVI	0.17
STAT-2	NDVI	0.25
STAT-2	SAVI	0.37
STAT-2	NDWI	-0.06
STAT-3	EVI	0.20
STAT-3	NDVI	0.28
STAT-3	SAVI	0.47
STAT-3	NDWI	-0.01
STAT-4	EVI	0.21
STAT-4	NDVI	0.37
STAT-4	SAVI	0.64
STAT-4	NDWI	0.15
STAT-5	EVI	0.21
STAT-5	NDVI	0.38
STAT-5	SAVI	0.52
STAT-5	NDWI	0.07
STAT-6	EVI	0.25
STAT-6	NDVI	0.42
STAT-6	SAVI	0.58
STAT-6	NDWI	0.12

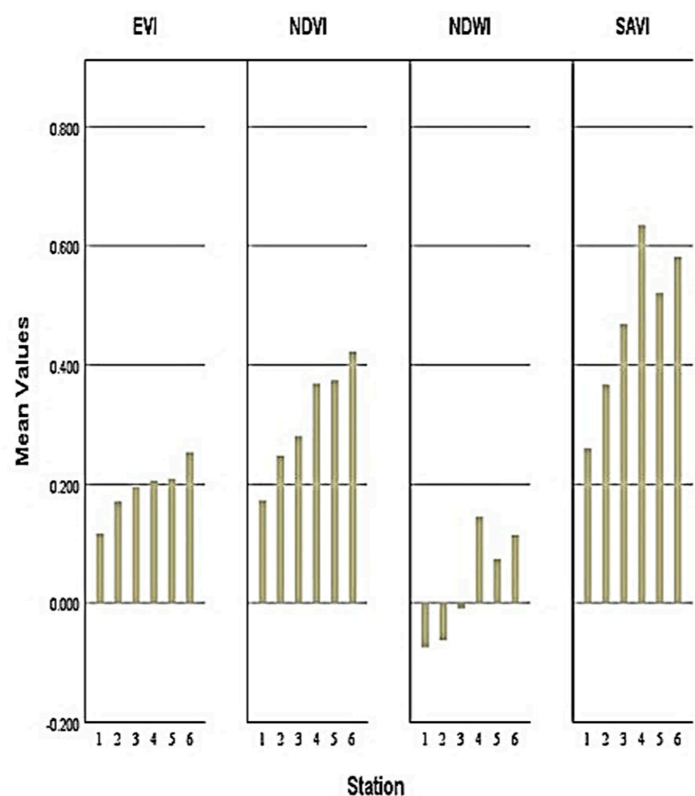


Figure 5. Mean of the vegetation indices EVI, NDVI, SAVI, and NDWI over the time series from 2001 to 2020.

The EVI values at each station vary from 0.12 to 0.25. Denser and healthier vegetation cover is generally indicated by higher EVI readings. The range of NDVI values is 0.17 to 0.42. NDVI counts the amount of vegetation and is a popular tool for evaluating the density and health of vegetation. The range of SAVI values is 0.26 to 0.64. With the adjustment of soil brightness, SAVI, a modified version of NDVI, offers a more realistic depiction of vegetation cover particularly in regions with elevated soil reflectance. The range of NDWI values is -0.07 to 0.15 . The water content of vegetation can be detected and tracked using NDWI. Delineating spatial configurations, temporal developments, and relationships between the health of the vegetation and environmental characteristics such as land use, land cover, and climate are made possible by this dataset.

The NDVI was measured between the years 2000 and 2005; it showed low values in the north and almost nonexistent values in the extreme north. On the other hand, it showed a strong presence that increased with distance to the extreme south (Figure 6).

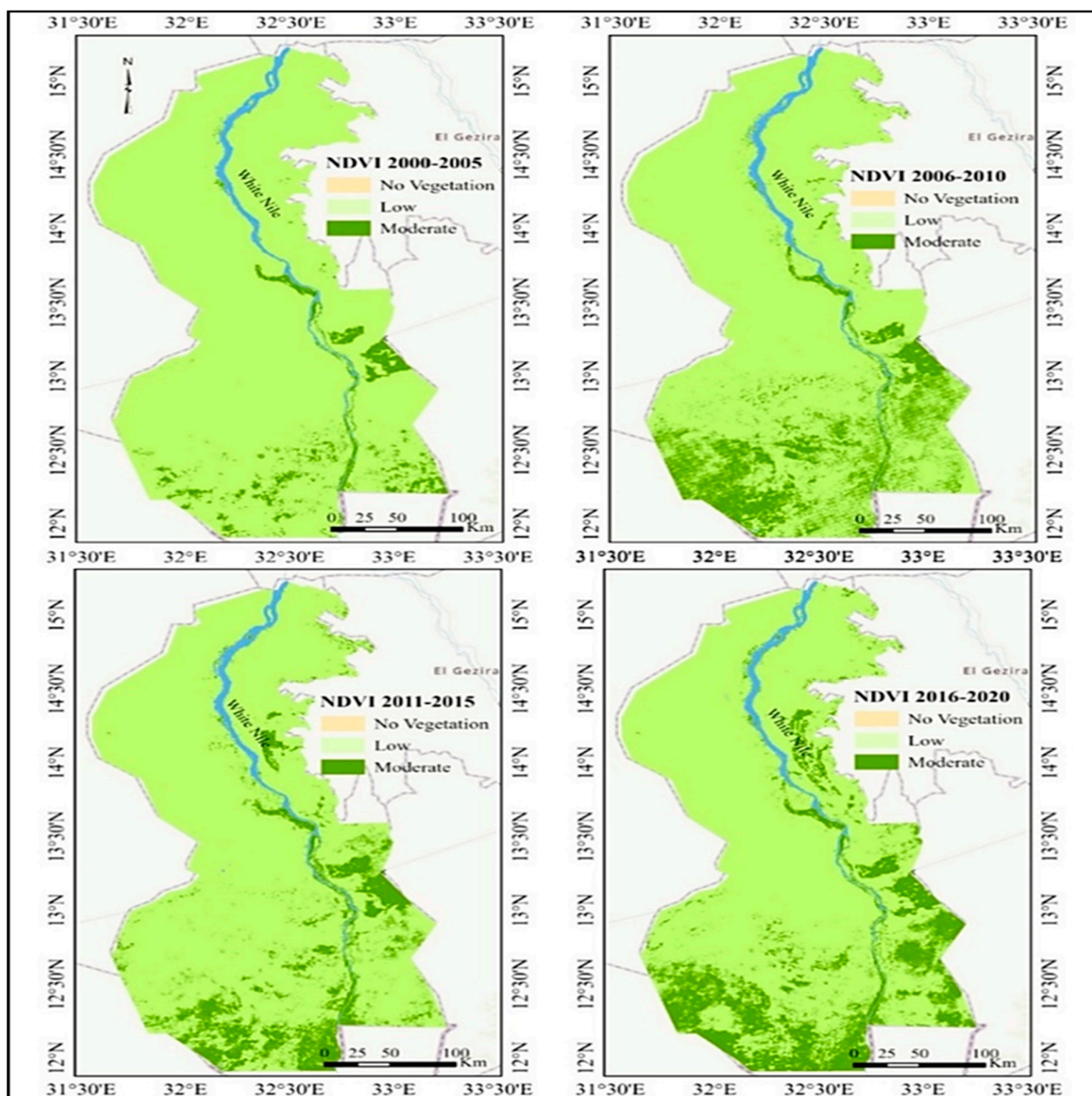


Figure 6. The difference in NDVI over the time series from 2001 to 2020 of White Nile State.

From 2006 to 2010, there was an increasing tendency in the normalized differential vegetation index towards the southern regions, with the far south experiencing the highest density. The normalized difference vegetation index decreased between 2011 and 2015, especially in the north, with nearly nonexistent levels in the far north; in contrast, vegetation density rose as we got closer to the southern regions. Regarding the time frame from 2016 to 2020, the normalized difference vegetation index was scarce, perhaps absent in the north, but it rose as we approached the south, resulting in higher vegetation densities, likely due to heavier precipitation (Figure 6). According to Balaghi, Tychon [61], NDVI data are especially thought to be most helpful in arid areas or where there are significant interannual fluctuations in the vegetation status. Therefore, when the yield forecasting model was being developed, it was anticipated that NDVI would offer more information.

The soil-adjusted vegetation index (SAVI) was measured between 2000 and 2005, with lower values in the northern parts and essentially no measurements in the northernmost sections (Figure 7).

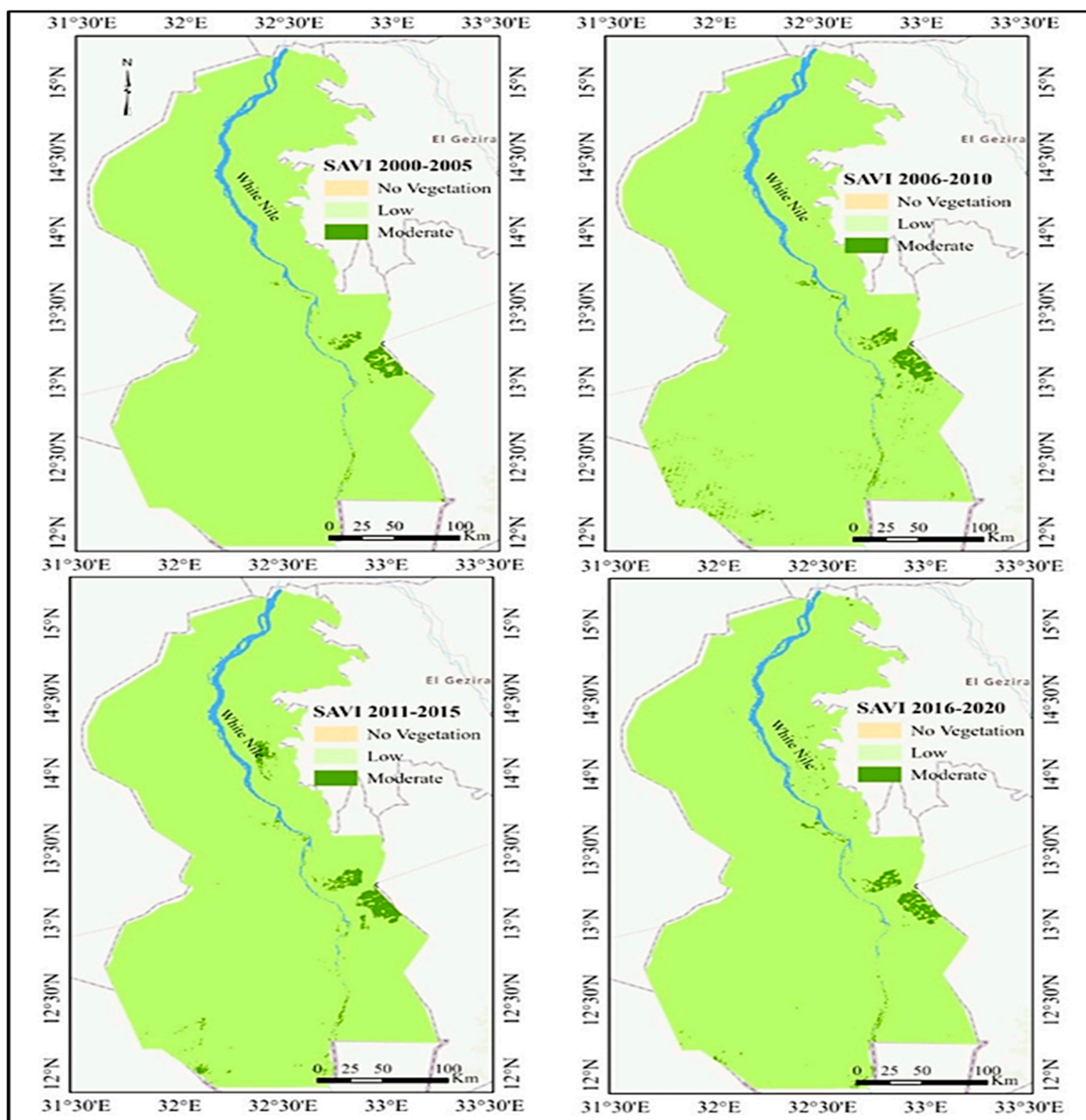


Figure 7. The difference in SAVI over the time series from 2001 to 2020 of White Nile State.

In contrast, the southern extremes of the White Nile district saw a significant increase in density. From 2006 to 2010, SAVI values decreased as one travelled towards northern territories that may have lacked SAVI presence, whereas SAVI levels gradually increased towards the southern regions, culminating in density at the farthest southern points (Figure 7). The period from 2011 to 2015 showed lower SAVI in the northern locations, almost nonexistent in the northernmost regions, and increasing vegetation density in the southern locations. For the years 2016 to 2020, there was a small amount of SAVI found, which may not have existed in the north. SAVI values increased in the south, especially in the extreme south, where they were linked to higher precipitation levels.

The normalized difference water index (NDWI) was observed during the time frame spanning from 2000 to 2005, exhibiting low values in the northern regions of the White Nile district and being almost nonexistent in the far north (Figure 8).

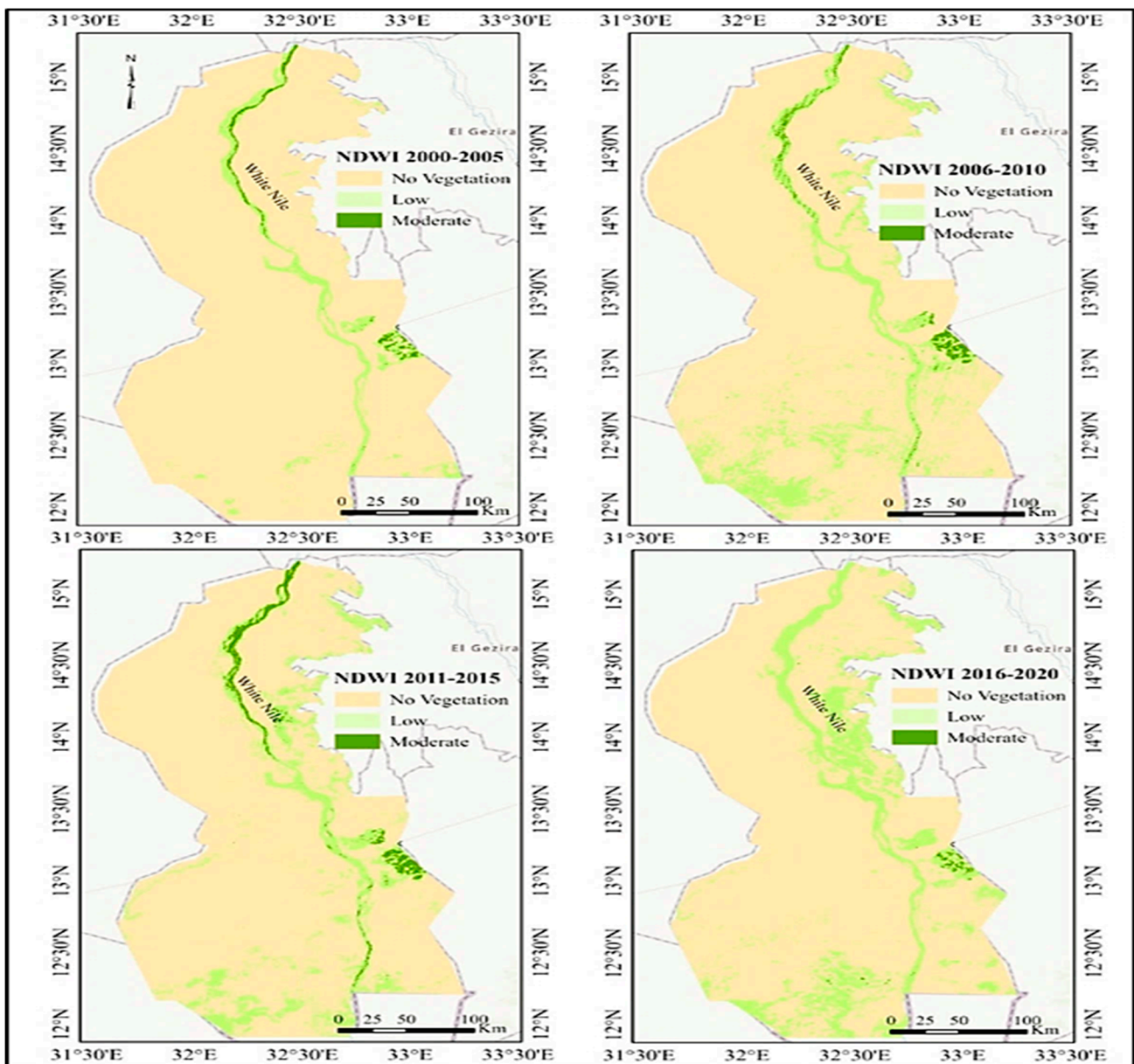


Figure 8. The difference in NDWI over the time series from 2001 to 2020 of White Nile State.

Conversely, a higher density of NDWI was noted in the southern regions, gradually increasing towards the far south. Between 2006 and 2010, a decline in the normalized difference water index (NDWI) was observed as one moves towards the northern areas. In

these regions, the NDWI might be absent, while its values rise as we move southwards, peaking in density in the far southern areas. From 2011 to 2015, a decrease in the NDWI was observed, particularly in the northern and far northern regions, with water density increasing as we approach the southern areas. As for the period spanning from 2016 to 2020, there were minimal quantities of the NDWI, possibly lacking in the northern regions, with an increase in NDWI values towards the south. The density of the NDWI peaks in the far south due to higher rainfall intensity (Figure 8).

The Enhanced Vegetation Index (EVI) had low values in the White Nile State’s northern regions from 2000 to 2005 but was higher in the southern regions. Between 2006 and 2010, EVI decreased moving north and increased moving south. From 2011 to 2015, EVI decreased in the north and increased in the south. From 2016 to 2020, EVI was minimal in the north and increased towards the south, peaking in the far south due to high rainfall (Figure 9).

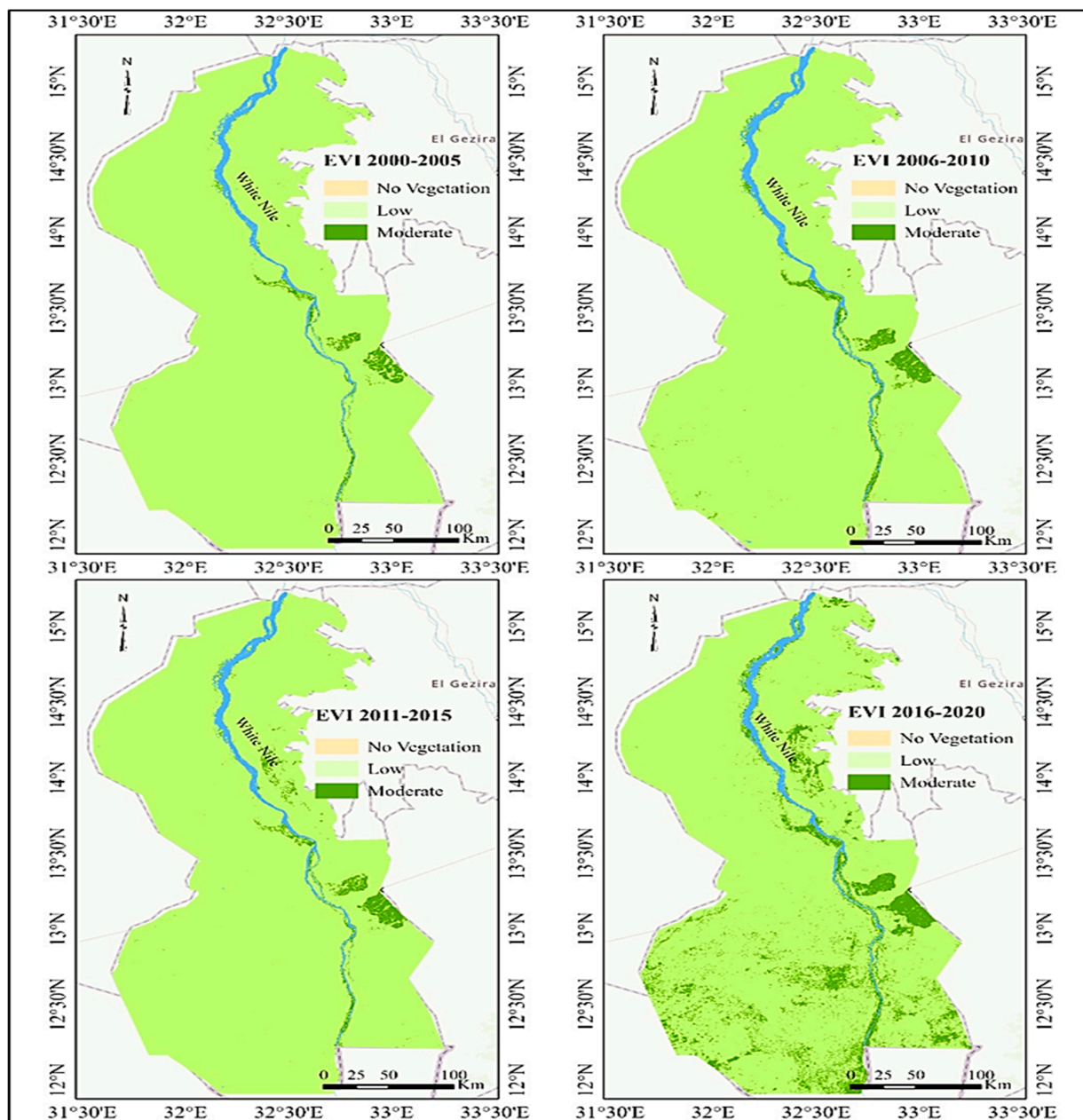


Figure 9. The difference in EVI over the time series from 2001 to 2020 of White Nile State.

Table 4 shows the areas that vary at regular intervals for every vegetation index. This draws attention to the variations in the planted cane areas between 2001 and 2020, which are a reflection of shifting climatic conditions, variations in rainfall, and the consequent availability of water needed for sugar cane irrigation. Since these regions were the least attainable during the period, we discovered that there was a significant fluctuation in the areas that were evident in the vegetation indices. The agricultural area declined between 2001 and 2005, then increased between 2005 and 2010; however, it was still greater than in the first era. From 2016 to 2020, the cultivated area increased once more.

Table 4. The changes in areas occur through periodic intervals for each vegetation index.

Vegetation Index	Classes	Area							
		(2000–2005)		(2006–2010)		(2011–2015)		(2016–2020)	
		(km ²)	(%)	(km ²)	(%)	(km ²)	(%)	(km ²)	(%)
NDVI	No Veg.	816.46	2.04	837.64	2.10	851.378	2.13	761.27	816.46
	Low	37,199.34	93.09	33,347.96	83.46	34,352.1	85.97	32,210.81	37,199.34
	Moderate	1942.90	4.86	5773.10	14.45	4755.19	11.90	6986.57	1942.90
SAVI	No Veg.	702.94	1.76	760.40	1.90	730.998	1.83	622.38	702.94
	Low	38,933.90	97.44	38,606.98	96.62	38,572.5	96.53	38,838.59	38,933.90
	Moderate	321.76	0.81	591.32	1.48	655.174	1.64	497.72	321.76
NDWI	No Veg.	39,111.52	97.88	34,207.52	85.61	34,056.7	85.23	33,498.87	39,111.52
	Low	320.63	0.80	5374.52	13.45	4874.58	12.20	6382.57	320.63
	Moderate	32.06	0.08	372.66	0.93	678.6	1.70	77.26	32.06
EVI	No Veg.	921.86	2.31	960.04	2.40	952.29	2.38	885.93	921.86
	Low	37,911.68	94.88	37,515.68	93.89	37,648.95	94.22	34,761.71	37,911.68
	Moderate	544.5	1.36	902.07	2.26	776.80	1.94	3730.40	544.5

3.3. Evaluation of the Machine Learning Models

Five input scenarios were produced by combining all 13 inputs, and the models now need to be trained on these. After being randomly assigned to the training and testing phases, the input scenarios were fed into the models. The models were put into practice, and the MBE, MAPE, and MARE criteria were used to assess the estimations of the models (Table 5). While there are clear distinctions between the models, there are also significant parallels concerning certain situations. The closer the MAPE, MBE, and MARE are to one, the better the model performance.

The models demonstrate good performance in GWFP estimation as the greatest MAPE value of 120.97 was attained under RF and Sc3, while the lowest value of -8.30 was noted under the SVR model and Sc3. The finest MBE under RF and Sc3 was $5.14 \text{ m}^3 \text{ ton}^{-1}$, the highest reported MBE under RF-SVR was $5.05 \text{ m}^3 \text{ ton}^{-1}$, and the lowest MBE under RF-SVR and Sc1 was 0.03. Table 5 shows that the maximum MARE values for RF and XGB with Sc3 were 59.23 and 57.82, respectively, while the lowest values were -4.05 and -3.99 for SVR with Sc2 and Sc1, respectively.

The relationship between the actual and predicted GWFP values is represented in Table 6 with a coefficient of determination (R^2) and the best-fitting equation for the relationship. Among the various equations tested, the linear equation yielded the highest R^2 value compared to exponential and logarithmic equations. The analysis revealed that the lowest R^2 values were observed for Sc3, which represented remote sensing indices (EVI, NDVI, SAVI, and NDWI), across all single and hybrid ML models. Conversely, the highest R^2 values were obtained for Sc1 (all parameters), Sc2 (climatic parameters), and Sc4 (Pe, Tmax, Tmin, and SA). Notably, the SVR model consistently achieved the highest R^2 values across all scenarios. Interestingly, the R^2 values for double hybrid models were higher than those for triple hybrid models, and the SVR model outperformed both double and triple hybrid models. The lowest R^2 value of 0.232 was observed for XGB and Sc3, while the highest R^2 value of 0.9846 was achieved for SVR and Sc4.

Table 5. Performance statistics of ML models applied to the five distinct climate and remote sensing variable scenarios.

Model	Index	Input Scenario				
		Sc1	Sc2	Sc3	Sc4	Sc5
RF	MBE (m ³ ton ⁻¹)	0.48	0.65	0.34	0.54	0.57
	MAPE	2.58	1.29	120.97	1.48	1.51
	MARE	1.26	0.63	59.28	0.73	0.74
XGB	MBE (m ³ ton ⁻¹)	1.01	0.96	-1.51	1.05	1.05
	MAPE	-6.28	-6.83	118.00	-8.05	-1.56
	MARE	-3.08	-3.35	57.82	-3.95	-0.76
SVR	MBE (m ³ ton ⁻¹)	1.10	1.09	5.14	0.87	1.13
	MAPE	-8.14	-8.30	98.31	-2.36	1.06
	MARE	-3.99	-4.07	48.17	-1.16	0.52
RF-XGB	MBE (m ³ ton ⁻¹)	0.98	1.17	1.87	0.75	1.09
	MAPE	-2.58	-1.53	92.66	3.78	-1.35
	MARE	-1.26	-0.75	45.40	1.85	-0.66
RF-SVR	MBE (m ³ ton ⁻¹)	0.03	-0.06	5.05	0.22	0.15
	MAPE	-1.83	-1.12	64.53	-0.72	0.25
	MARE	-0.90	-0.55	31.62	-0.35	0.12
XGB-SVR	MBE (m ³ ton ⁻¹)	0.33	0.48	2.86	0.06	0.26
	MAPE	0.06	-1.13	60.44	5.48	0.87
	MARE	0.03	-0.55	29.61	2.68	0.43
RF-XGB-SVR	MBE(m ³ ton ⁻¹)	0.97	1.32	2.44	1.25	0.88
	MAPE	3.61	-0.47	99.89	-0.71	5.16
	MARE	1.77	-0.23	48.94	-0.35	2.53

Table 6. The coefficient of determination between actual and predicted GWFP values and the best-fitting equation.

Model	Input Scenario	R ²	Fitting Equation
RF	Sc1	0.9662	y = 0.8619x + 2.581
	Sc2	0.9621	y = 0.8532x + 2.604
	Sc3	0.2407	y = 0.222x + 16.925
	Sc4	0.9680	y = 0.8633x + 2.49
	Sc5	0.9680	y = 0.8633x + 2.49
XGB	Sc1	0.9671	y = 0.8757x + 1.7511
	Sc2	0.9688	y = 0.8759x + 1.7894
	Sc3	0.2322	y = 0.3305x + 16.37
	Sc4	0.9618	y = 0.874x + 1.7483
	Sc5	0.9499	y = 0.8541x + 2.1941
SVR	Sc1	0.9816	y = 0.8826x + 1.503
	Sc2	0.9802	y = 0.8798x + 1.5772
	Sc3	0.2390	y = 0.0959x + 14.929
	Sc4	0.9846	y = 0.8807x + 1.7732
	Sc5	0.9730	y = 0.8647x + 1.8714
RF-XGB	Sc1	0.9727	y = 0.8873x + 1.8993
	Sc2	0.9625	y = 0.8943x + 1.9451
	Sc3	0.2528	y = 0.1897x + 15.18
	Sc4	0.9683	y = 0.8743x + 2.2113
	Sc5	0.9601	y = 0.8829x + 2.1468

Table 6. Cont.

Model	Input Scenario	R ²	Fitting Equation
RF-SVR	Sc1	0.971	$y = 0.9155x + 1.8096$
	Sc2	0.9668	$y = 0.9155x + 1.8931$
	Sc3	0.2811	$y = 0.1536x + 13.344$
	Sc4	0.9714	$y = 0.8969x + 2.022$
	Sc5	0.9692	$y = 0.9013x + 1.9928$
XGB-SVR	Sc1	0.9781	$y = 0.9203x + 1.3998$
	Sc2	0.9809	$y = 0.9198x + 1.2626$
	Sc3	0.3409	$y = 0.2536x + 13.368$
	Sc4	0.9782	$y = 0.9094x + 1.907$
	Sc5	0.9683	$y = 0.9167x + 1.5544$
RF-XGB-SVR	Sc1	0.9593	$y = 0.7911x + 3.4766$
	Sc2	0.9555	$y = 0.7801x + 3.3624$
	Sc3	0.4960	$y = 0.3492x + 11.399$
	Sc4	0.9549	$y = 0.785x + 3.3223$
	Sc5	0.9487	$y = 0.777x + 3.8615$

During the test period, bar charts were utilized to compare the ML models, whether single or hybrid, and scenarios based on the CC, NSE, MAE, and Tstat criteria. These comparisons were made separately for each scenario. The comparison charts, also known as combo bar charts, revealed significant variations in the values of CC. The analysis indicated that the highest CC value of 0.99 was observed under Sc1 and Sc4 for SVR, RF-XGB, RF-SVR, XGB-SVR, and RF-XGB-SVR, surpassing both XGB and RF with a CC value of 0.98. Following closely were Sc2, Sc4, and Sc5, with differences ranging from 0.01 to 0.02, making all scenarios, except Sc3, relatively similar. However, in the case model, the lowest CC value of 0.48 was achieved under Sc3, which took into consideration NDVI, SAVI, and NDWI (Figure 10).

On the contrary, the NSE exhibited a similar pattern of data to CC, with slight discrepancies in values. The highest NSE of 0.98 was observed under Sc2 (climatic parameters) and XGB-SVR. The variance between ML models, whether single or hybrid, ranged from 0.01 to 0.03 under the same conditions, except for Sc2, where the difference between XGB-SVR and XGB was 0.05 higher for hybrid models. The lowest NSE value of 0.09 was recorded with SVR and Sc3. NSE values for Sc1, Sc2, Sc4, and Sc5 were considered very good across all models, while for Sc3, the data from all single and hybrid models were classified as satisfactory to unsatisfactory (Figure 10).

Similarly, the MAE displayed a similar trend to MSE and CC, with the highest MAE of 1.00 under Sc1 (all parameters) and RF-XGB-SVR. The lowest MAE value of 0.18 was recorded with SVR and Sc3. The lowest-performing scenario across all ML models, whether single or hybrid, was Sc3. The highest Tstat value of 2.26 was observed under Sc1 (all parameters) and the SVR model. Conversely, the lowest Tstat value of 0.06 was recorded with RF-SVR and Sc1 (Figure 10).

Scenarios Sc4 and Sc5 were found to be favourable for all models compared to Sc1 and Sc2 due to having fewer parameters, high accuracy, and low errors, which closely resembled Sc1 and Sc2. Therefore, if remote sensing images are available, weather stations are not updated, and only P_{eff} , T_{max} , T_{min} , and SA can be measured, it would be satisfactory to achieve accurate predictions for GWFP, especially in low-income countries.

The performance of machine learning models was assessed across five different climate and remote sensing scenarios. However, a distinct evaluation of climate change (CC) and scenario criteria was obtained for the remaining scenarios. Various metrics were employed to gauge the accuracy of the scenarios, aiming to extract more relevant insights from the time series data. Scenario model accuracy refers to the percentage of accurate predictions for the test data, indicating how closely the actual values align with the predicted values obtained through the specific scenario and prediction model. Achieving accuracy levels

above 90% is considered satisfactory. In some scenarios and models, values exceeding 1 were noted, suggesting that the prediction outcomes might be influenced by random errors, which could be attributed to measurement errors rather than systematic factors. Particularly, RF and XGB models showed overestimated scenario values, which could be attributed to random errors.

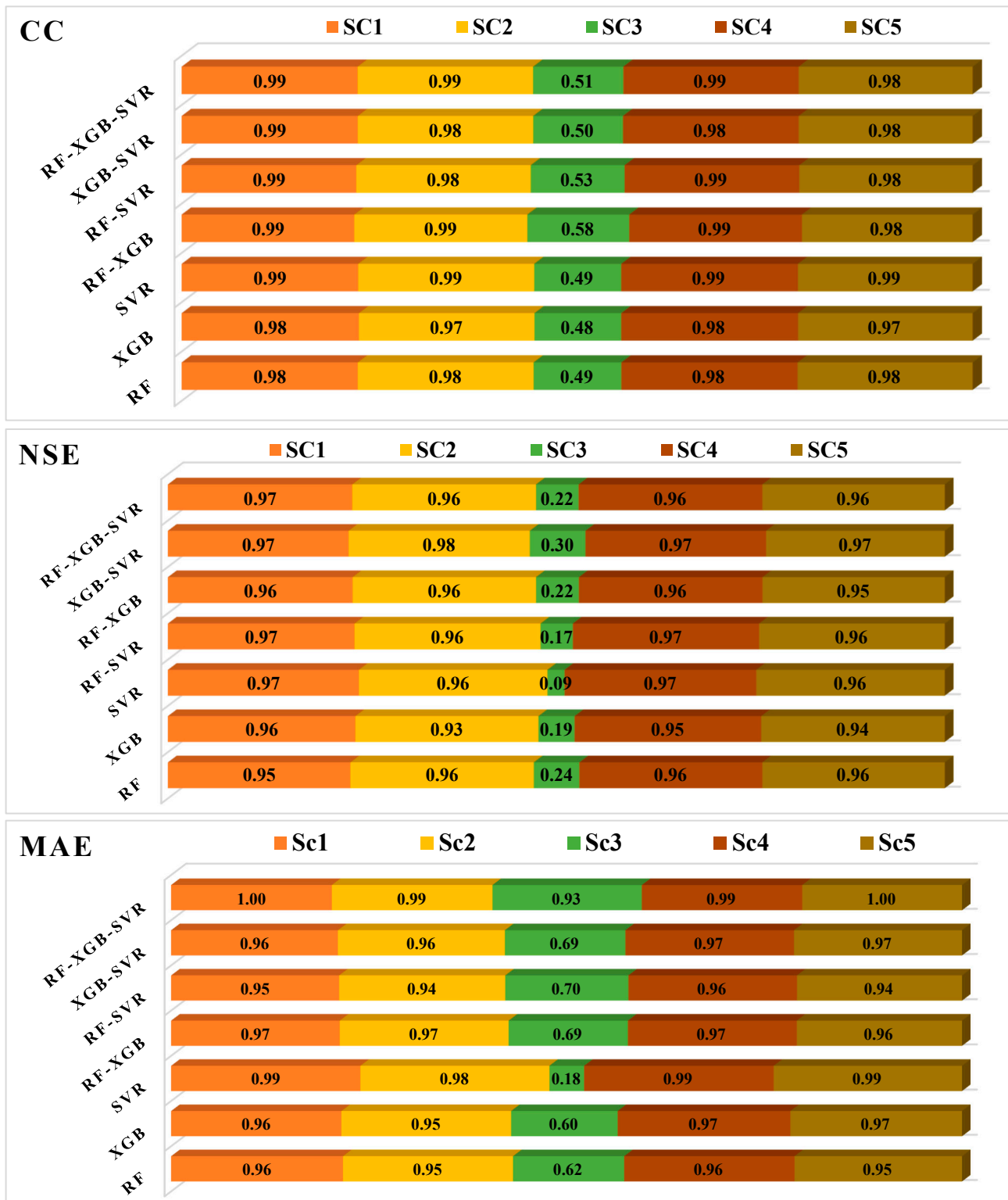


Figure 10. Cont.

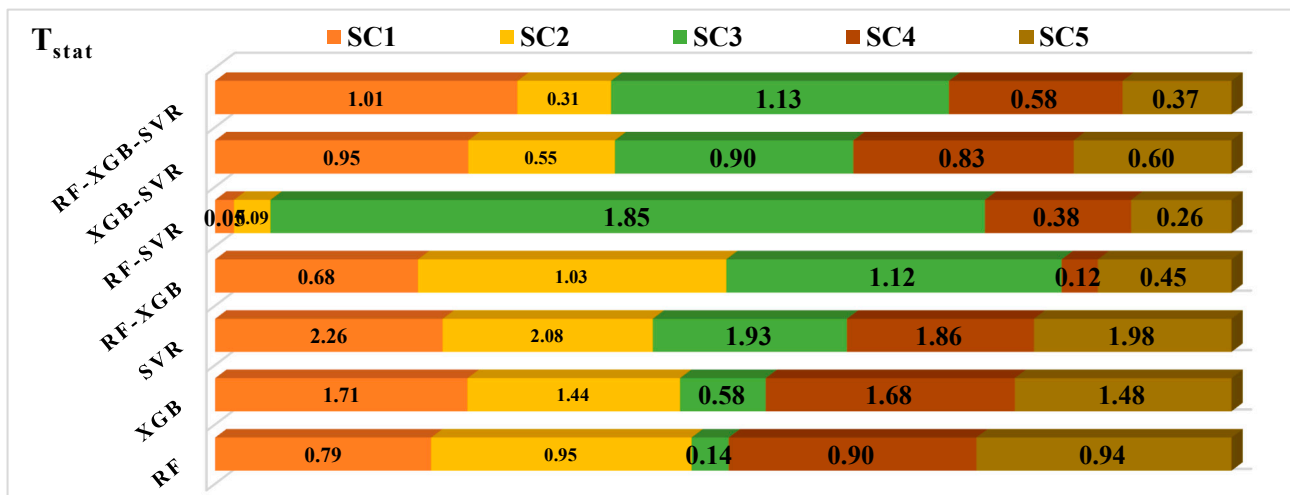


Figure 10. Bar charts to compare the models in each scenario separately, based on the CC, NSE, MAE, and T_{stat} .

Radar charts were utilized to illustrate the RMSE of the GWFP for three different models, both single and hybrid, across five scenarios (Figure 11). The highest RMSE value in Sc1 was recorded with the RF model at 4.28, followed by XGB at 4.21, while the lowest RMSE values were seen with the XGB-SVR hybrid models at 3.42. Moving on to scenario 2, the highest and lowest RMSE values were observed with XGB at 5.19 and XGB-SVR at 3.28, respectively. In scenario 3, higher RMSE values were evident compared to other scenarios, with the highest values under RF and RF-XGB-SVR at 19.55 and 19.24, respectively, and the lowest RMSE value under RF at 17.5. For scenario 4, the highest and lowest RMSE values were found with XGB at 4.44 and SVR at 3.37, respectively. Finally, in scenario 5, the highest and lowest RMSE values were registered with XGB at 5.01 and XGB-SVR at 3.99, respectively (Figure 11).

3.4. Accuracy and Uncertainty of the Models

Figure 12 illustrates a comparison of combo graphs between scenarios and machine learning models based on uncertainty criteria (U_{95}) and accuracy. The analysis reveals significant differentiation in the values of accuracy and U_{95} . The uncertainty of scenarios used in modelling was assessed based on the limits of the 95% confidence interval, signifying the likelihood of obtaining results close to the expected value within the defined uncertainty range. It was observed that the highest U_{95} was associated with Sc3 scenarios, considering temperature, wind speed, and crop factor, leading to the lowest accuracy. RF-SVR for Sc3 showed the highest U_{95} at 53.56, while XGB-SVR for Sc2 had the lowest at 9.10, with uncertainty values ranging from 9.10 to 14.30 for all scenarios except Sc3. Notably, the lowest uncertainty values were recorded for Sc1 across all models except XGB-SVR, where Sc2 surpassed it. The accuracy analysis indicated the highest accuracy levels for Sc1 and Sc4 scenarios in SVR and XGB models, respectively, while the lowest accuracy values were found under Sc3, particularly under RF. The accuracy values for Sc1, Sc2, Sc4, and Sc5 were above or equal to 0.95, displaying excellent performance in predicting GWFP.

On the contrary, the analysis of accuracy indicated that the Sc1 and Sc4 scenarios achieved the highest accuracy levels of 1.08 in the SVR and XGB models, respectively. Conversely, the lowest accuracy values were observed in Sc3, with RF performing the worst by -0.21 . It was challenging to differentiate between scenarios Sc1, Sc2, Sc4, and Sc5, as their accuracy values were all above or equal to 0.95, with differences not exceeding 1–2%. Despite this, their predictive performance for GWFP was exceptional (Figure 12).

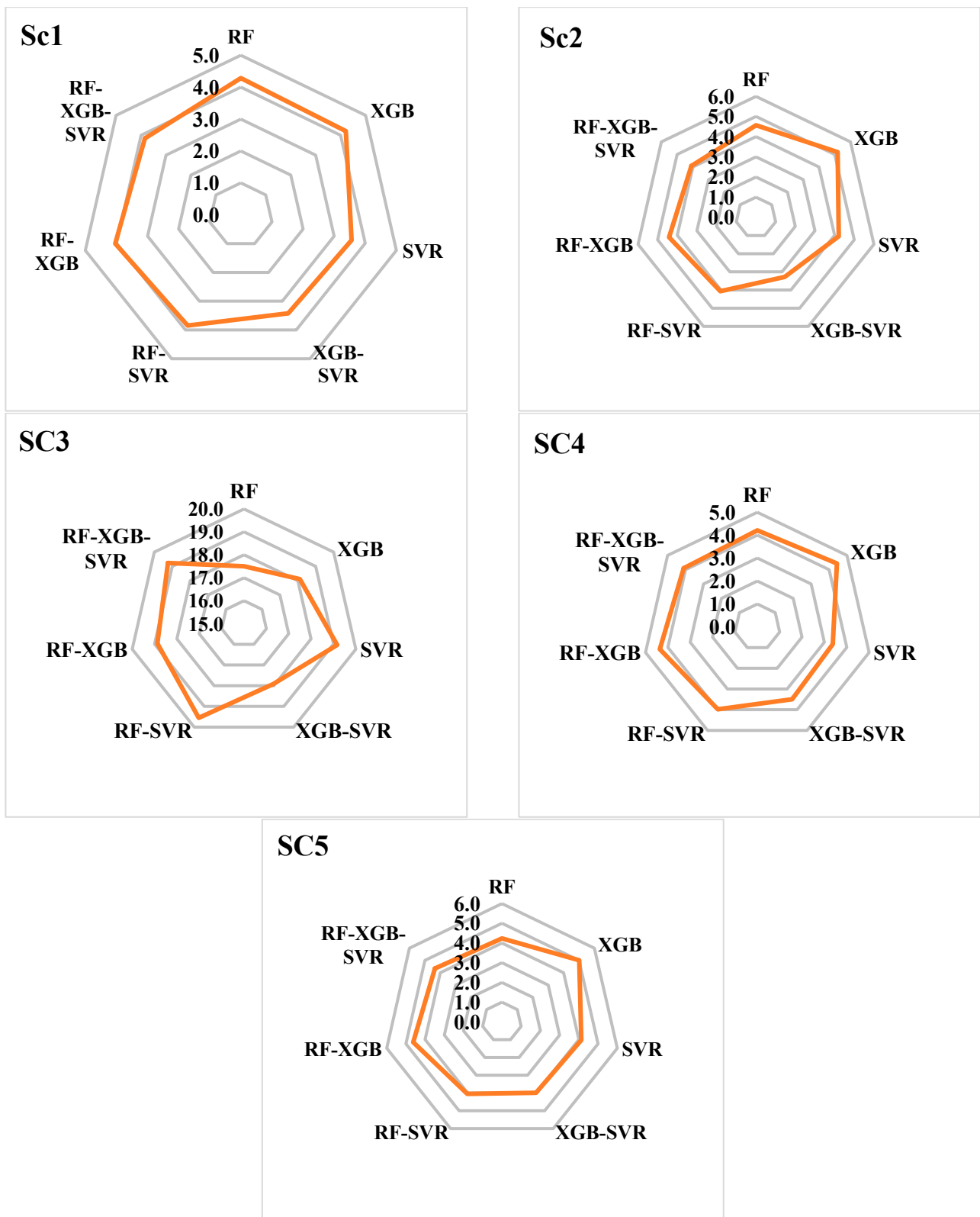


Figure 11. Radar charts to compare the models in each scenario separately, based on the RMSE criterion.

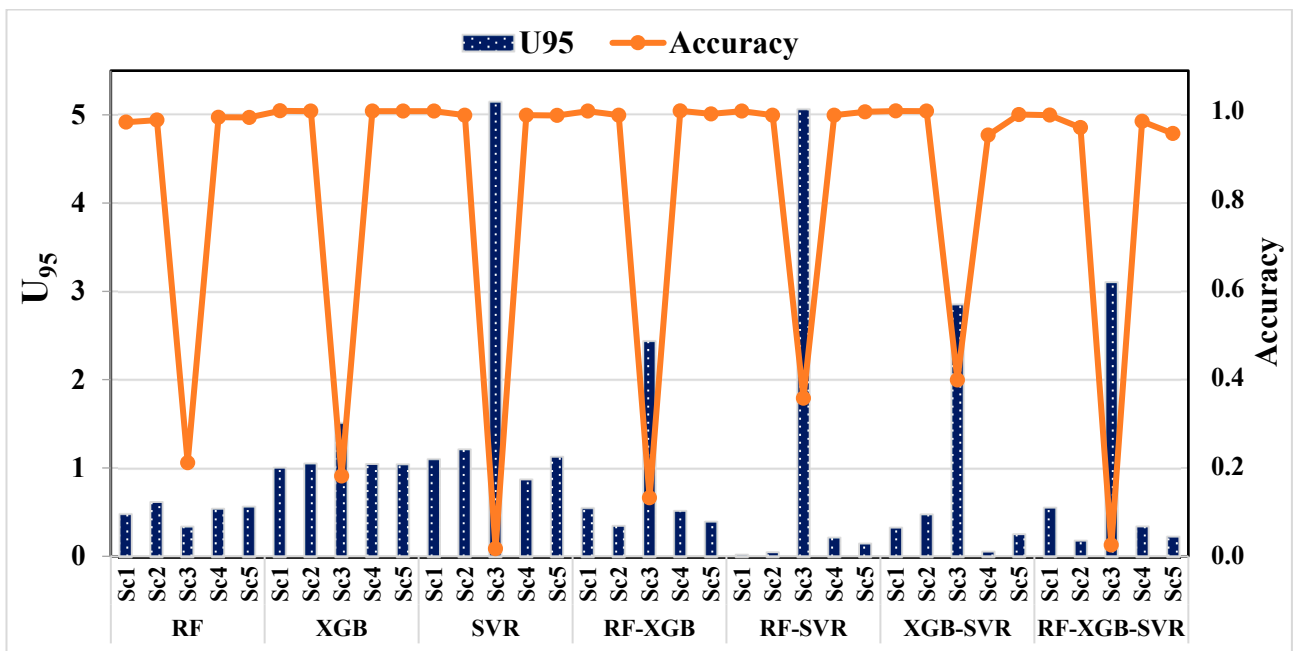


Figure 12. Combo graphs for comparison between the scenarios and ML models based on the criteria U_{95} and accuracy.

3.5. Comparison of the Machine Learning Models

The Violin plots depict defect distribution based on four key values: first quartile (Q1), third quartile (Q3), interquartile range (IQR), and the median. The XGB-SVR hybrid models with Sc2 exhibited the lowest error IQR at 0.359, whereas the XGB model with Sc1 had the highest IQR at 2.72. A lower IQR indicates an error distribution close to zero, with the median line representing a normal error distribution. For GWFP prediction, the most effective models were XGB-SVR with Sc2, followed by RF-XGB-SVR with Sc1, and then RF-SVR with Sc1 (Figure 13).

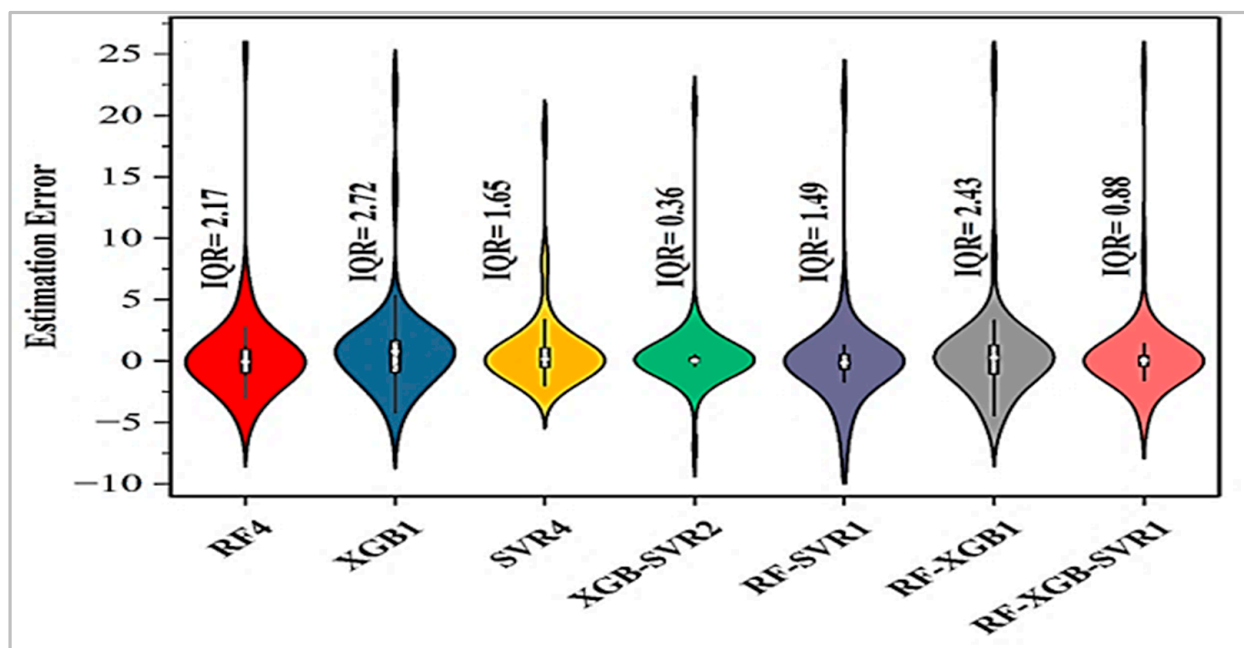


Figure 13. Violin plot for error distribution of models in different scenarios.

Figure 14 offers a visual representation of the importance of 13 input variables and their respective contributions to GWFP. These contributions are scaled between 0 and 1, with a value of 1 denoting the highest impact on the target variable.

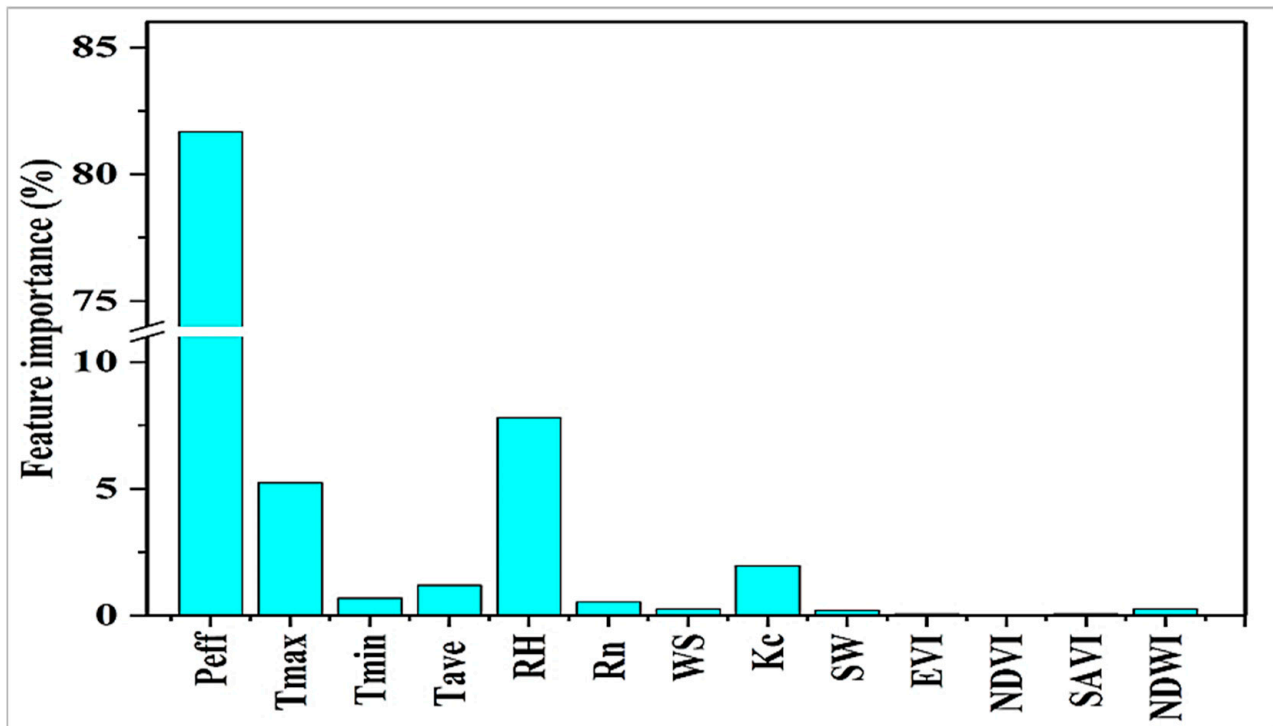


Figure 14. Relative contributions of 13 input variables to green water footprint.

Effective precipitation emerges as the most influential variable, accounting for 81.67% of the impact on GWFP. Following effective precipitation, relative humidity (RH) had a 7.5% impact, maximum temperature was 5.24%, crop coefficient k_c was 2.5%, and minimum temperature was 0.6956%. EVI and NDVI also displayed significant impacts on the green water footprint.

3.6. Response of GWFP to Climate, Crop, and Remote Sensing Parameters

The correlation coefficients between GWFP and the following parameters (effective precipitation (P_{eff}), minimum temperature (T_{min}), relative humidity (RH), solar radiation (R_n), EVI, NDVI, and NDWI) are positive (Figure 15). The highest correlation coefficient was 0.99 between GWFP and P_{eff} and 0.34 between GWFP and RH, whereas the fraction of the NDVI had a very low positive correlation with GWFP at 0.0086. Since the growing season occurs in the summer in Sudan, which is situated in a dry region and relies mainly on rainfall for irrigation, the effective precipitation showed a strong correlation with GWFP at 0.99. Factors such as maximum temperature (T_{max}), wind speed (WS), ETC, yield, sawn area, and SAVI exhibited a negative correlation with GWFP. Specifically, the maximum temperature displayed the highest negative correlation coefficient of -0.24 with GWFP. Through the examination of correlation coefficients, it was observed that climatic parameters, followed by remote sensing data, and ultimately crop parameters, had the most significant influence on GWFP.

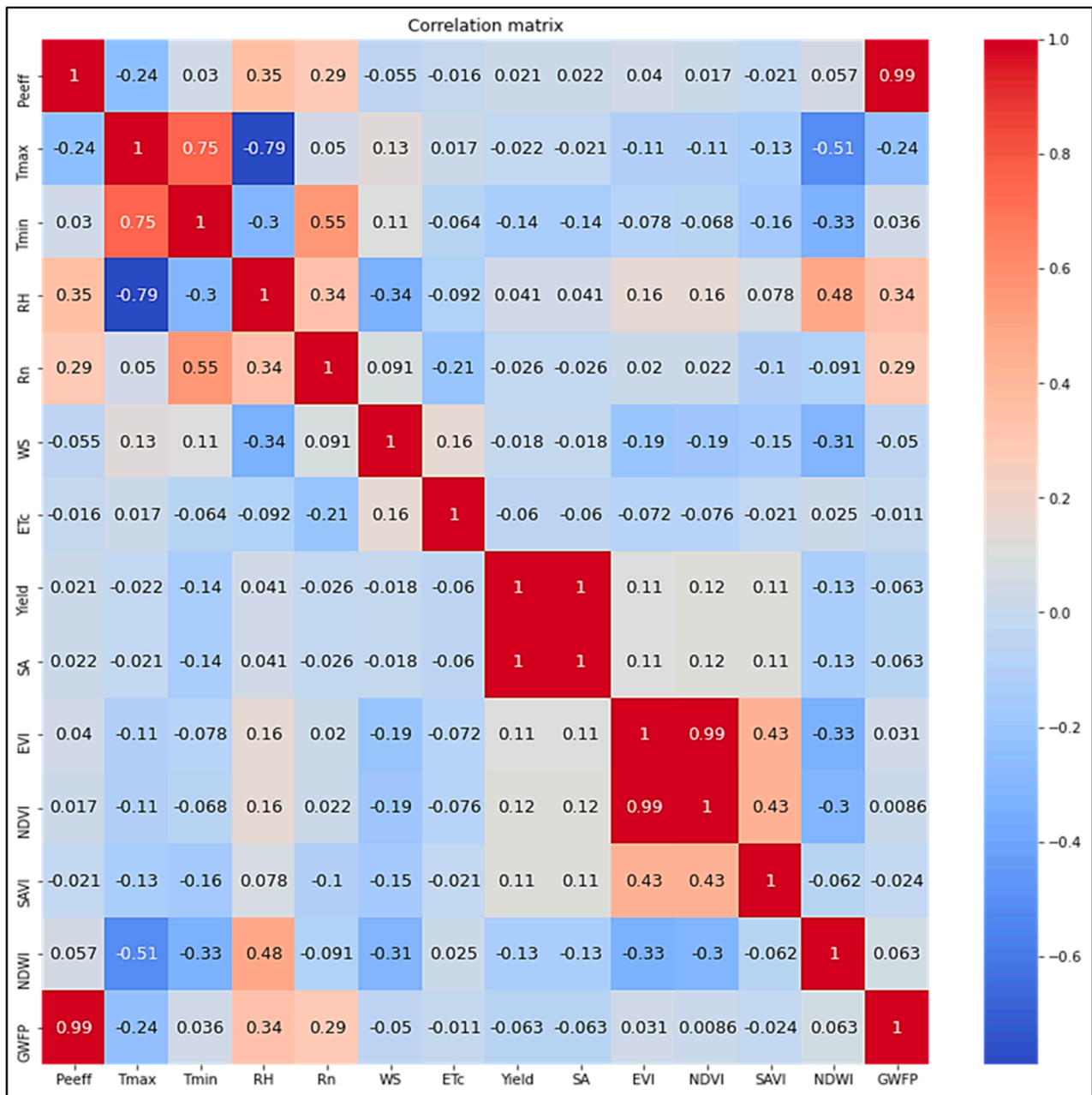


Figure 15. Correlation matrix between effective precipitation, maximum temperature, minimum temperature, relative humidity, solar radiation, wind speed, crop evapotranspiration, yield, sown area, EVI, NDVI, SAVI, NDWI, and green water footprint.

4. Discussion

The inadequate efficacy of remote sensing indices in forecasting the green water footprint can be attributed to the reliance on climatic data for the computation of green water footprints, which principally involves the estimation of evapotranspiration during the vegetative period, alongside another climatic factor, namely effective rainfall in conjunction with agricultural productivity. Consequently, models incorporating climatic data demonstrate a robust and efficacious capacity in predicting the green water footprints associated with sugarcane. Conversely, remote sensing indices exhibited a lack of correlation with methodologies employed for the estimation of water footprints.

The diminished performance of Sc3, representing remote sensing indices, may be improved by augmenting predictive accuracy through advanced remote sensing techniques. This enhancement can be realized firstly by utilizing higher-resolution remote sensing

datasets, such as Sentinel 1 and 2, which are anticipated to better capture the nuanced variability of fields in contrast to lower-resolution datasets. Secondly, an increase in temporal resolution may elevate the frequency of data acquisition, thus facilitating more accurate monitoring of phenological changes, particularly during critical developmental stages of wheat. Finally, the incorporation of additional remote sensing indices, such as the Normalized Difference Temperature Index (NDTI), could be advantageous in clarifying plant stress, whereas Gross Primary Productivity (GPP) provides a direct link to crop water consumption, and Soil Moisture along with Land Surface Temperature (LST) aids in comprehending the hydrological processes relevant to sugarcane agriculture.

Tao, Zhang [62] predicted the green and blue water footprints (GWFPs and BWFPs) of cassava in Nanning, Guangxi, China, using an Artificial Intelligence–Seasonal ARIMA (AI-SARIMA) integrated model. In addition to the SARIMA model for time series forecasting, they used three supervised learning algorithms: random forests (RFs), support vector machines (SVMs), and artificial neural networks (ANNs). To simulate WFs under various climate scenarios, meteorological data from 1994 to 2019—including T_{min}, T_{max}, P, SH, WS, and H—were employed. According to the results, the optimum ANN for estimating BWFP was an ANN with hidden layers (8, 6) and input variables T_{max}, T_{min}, P, K_c, SH, and H; for estimating GWFP, an ANN with hidden layers (7, 5) and input variables T_{max}, T_{min}, P, WS, and SH was ideal. These models had coefficients of determination that were almost one and exceptional accuracy. Our results are in line with those of Tao, Zhang [62], where the research results revealed that the single model followed by hybrid models obtained the highest R² where the SVR and Sc2 had a coefficient of determination of 0.9846 followed by the XGB-SVR hybrid model and Sc2 with R² value of 0.9809, while the lowest R² value was detected under XGB and Sc3 (remote sensing indices) at 0.2322.

In the same context, Mokhtar, El-Ssawy [63] investigated lettuce yield (fresh weight) prediction using four machine learning models, namely, support vector regressor (SVR), extreme gradient boosting (XGB), random forest (RF), and deep neural network (DNN). Three scenarios consisting of the combinations of input variables (i.e., leaf number, water consumption, dry weight, stem length, and stem diameter) were assessed. The XGB model with scenario 3 (all input variables) yielded the lowest root mean square error (RMSE) of 8.88 g followed by SVR with the same scenario that achieved 9.55 g, and the highest result was by RF with scenario 1 (i.e., leaf number and water consumption) that achieved 12.89 g. All model scenarios with scatter index (SI) values less than 0.1 were classified as excellent in predicting fresh lettuce yield. Based on all of the performance statistics, the two best models were SVR with scenario 3 and DNN with scenario 2 (i.e., leaf number, water consumption, and dry weight). Our research outcomes disagree with Mokhtar, El-Ssawy [63], where the highest RMSE values were obtained under single and hybrid ML models with RF and RF-XGB-SVR at 19.55 and 19.24 and Sc3 (remote sensing indices), while the lowest RMSE value was detected under XGB-SVR hybrid model and Sc2 (P_{eff}, T_{max}, T_{min}, and SA) at 3.28.

In the same context, which also supports the outputs of individual machine learning programmes, Abdel-Hameed, Abuarab [64] developed and compared four machine learning models—SVR, RF, XGB, and ANN—over three potato governorates (Al-Gharbia, Al-Dakahlia, and Al-Beheira) in the Nile Delta of Egypt to select the best model in the best combination of climate input variables to predict potato BWFP during 1990–2016. The available variables used were T_{max}, T_{min}, T_{ave}, WS, RH, P, VPD, SR, SA, and K_c. Six scenarios (Sc1–Sc6) of input variables were used. The findings indicated that Sc5, utilizing the XGB and ANN models, exhibited the most promising outcomes in predicting BWFP in a dry region by considering vapour pressure deficit, precipitation, solar radiation, and crop coefficient data, followed by Sc1 (incorporating all parameters). These developed models yielded notably superior results, offering valuable insights for water management and development planning decisions. The research results agree with a part of the results of Abdel-Hameed, Abuarab [64], which deals with the superiority of the single model over hybrid ML models where the single model achieved the lowest RMSE and highest R²

values followed by hybrid models, but they disagree with the part of best scenario where Sc4 (Pe, Tmax, Tmin, and SA) achieved the highest R^2 value with 0.98 like the Sc2 (all climatic variables). This supports the use of climate indicator monitoring stations with the least possible number of indicators and thus at the lowest price, which benefits developing countries with limited financial capabilities.

Ge, Zhao [65] employed the XGBoost regression (XGBR) model to estimate ET over three years (2019–2021), focusing on the impact of various meteorological factors on ET. Their study utilized a greenhouse drip-irrigated tomato crop ET prediction model (XGBR-ET) based on XGBoost regression, comparing it with seven other regression models. The importance of meteorological factors on XGBR-ET was ranked as follows: $R_n > RH > RH_{min} > T_{max} > RH_{max} > T_{min} > T_a > VPD$. Performance evaluation metrics R^2 , RMSE, and MAE were reported as 0.981, 0.163, and 0.132, respectively. Our results are not in line with those of Ge, Zhao [65], where the feature importance ranked as follows: $Pe_{eff} > RH > T_{max} > K_c > T_{ave} > T_{min} > R_n > WS > SA$. This is related to the fact that GWFP mainly depends on Pe_{eff} , which recorded the highest value in feature importance with 81.67%, followed by relative humidity (RH) with a 7.5% impact.

To optimize the SVR model, Elbeltagi et al. [66] used particle swarm optimization (PSO), RF, and SVR. Using various input meteorological variables, the hybrid RF–SVR–PSO model was assessed against a standalone SVR model, a back-propagation neural network (BPNN) model, and an RF model. All models were used to estimate ET_c ; the best model was SVR, with an R^2 of 0.97, followed by the hybrid RF–SVR–PSO model, with an R^2 of 0.975. The investigation of the GWFB's R^2 showed that the SVR and Sc2 had the highest R^2 at 0.9846 followed by the XGB-SVR hybrid model and Sc2 with an R^2 value of 0.9809, while the lowest R^2 value was detected under XGB and Sc3 (remote sensing indices) at 0.2322. This is strongly consistent with the research findings, as the analysis showed that the single model followed by hybrid models obtained the highest R^2 values.

In a research endeavour conducted by Wu et al. [67], the amalgamation of the Extreme Learning Machine (ELM) framework with two innovative meta-heuristic methodologies, specifically the Whale Optimization Algorithm (WOA) and the Flower Pollination Algorithm (FPA), was investigated for the forecasting of monthly pan evaporation (E_p). Hybrid models incorporating WOA-ELM and FPA-ELM were formulated to estimate monthly E_p within the Poyang Lake Basin located in Southern China. These hybrid models were evaluated against the differential evolution algorithm–optimized ELM (DEELM), the enhanced M5 model tree (M5P), and artificial neural network (ANN) frameworks. The findings revealed that the FPA-ELM model exhibited the highest predictive accuracy across all four monitoring stations, with the WOA-ELM model following closely behind, both outperforming conventional models. The application of heuristic algorithms, particularly the FPA, was strongly advocated for the enhancement of the efficacy of independent machine learning frameworks. The outcomes corroborated the results concerning RMSE, a pivotal metric for evaluating model accuracy in forecasting the GWFP, as our investigation illustrated that the hybrid models produced minimal values, thereby endorsing the efficacy of hybridization in diminishing errors and augmenting result precision (Figure 11).

Azzam, Zhang [68] utilized artificial neural networks (ANNs), support vector machines (SVMs), random forests (RFs), and k-nearest neighbours (KNNs) in their investigation to forecast green water evapotranspiration (GWET) and blue water evapotranspiration (BWET). Among the employed models, the random forest (RF) model demonstrated superior efficacy in estimating BWET, achieving a coefficient of determination (R^2) of 0.96, a mean inter-annual (MIA) of 0.91, a root mean square error (RMSE) of 10.77 mm month⁻¹, a Nash–Sutcliffe efficiency (NSE) of 0.92, and a mean absolute error (MAE) of 6.84 mm month⁻¹. Furthermore, the RF model, except for the Pre variable, exhibited satisfactory simulation outcomes ($0.3 \leq NSE < 0.6$), whereas all alternative machine learning algorithms displayed inadequate simulation results ($NSE < 0.3$). When evaluating the outcomes for predicted wheat BWFP, it was noted that the RF model in conjunction with Sc2 (climatic parameters) yielded the lowest NSE value of 0.05, which was regarded as inadequate. In

contrast, the highest NSE of 0.98 was recorded under Sc2 (climatic parameters), and the XGB-SVR hybrid model was categorized as perfect, while the lowest NSE of 0.09 was noted with SVR and Sc3, which was classified as inadequate. The NSE values for Sc1, Sc2, Sc4, and Sc5 were assessed as very good across all models, while the data for Sc3 from both single and hybrid models were categorized as satisfactory to unsatisfactory.

Conversely, the results of the research are inconsistent with the investigation carried out by Elhussiny, Hassan [69], who employed three distinct machine learning algorithms, specifically random forest (RF), extreme gradient boosting (XGB), and the hybrid random forest–extreme gradient boosting (XGB-RF), to predict the uniformity of water distribution from fixed set sprinklers. This analysis was predicated on several variables including operating pressure, sprinkler elevations, discharge rates, nozzle diameters, wind velocity, and humidity levels, along with maximum and minimum temperature readings. The findings revealed that the peak R^2 coefficients were 0.796, 0.825, and 0.929 for RF, XGB, and XGB-RF, respectively, in the initial context for CU. Likewise, for distribution uniformity (DU), the highest R^2 coefficients were recorded at 0.701, 0.479, and 0.826 for RF, XGB, and XGB-RF within the same context. It was noted that the hybrid XGB-RF model improved the R^2 coefficient by 10–13% relative to the standalone XGB and RF models. In contrast, our study indicated that the SVR model consistently achieved a superior R^2 coefficient across all contexts, surpassing both dual and triple hybrid models (Figure 9).

Hou, Yin [54] assessed the daily crop evapotranspiration (ET_c) of spring maize utilizing support vector regression (SVR). To determine the optimal input variables for the SVR framework, random forest (RF) was employed as a data preprocessing method. The SVR model underwent optimization through the application of particle swarm optimization (PSO). The efficacy of the innovative hybrid RF-SVR-PSO model was evaluated against a standalone SVR model, a back-propagation neural network (BPNN) model, and an RF model by incorporating various meteorological input variables. The Penman–Monteith equation was applied to derive the ET_c values, which served as a reference standard against the values estimated by the models. The findings indicated that the hybrid RF-SVR-PSO model exhibited superior performance in estimating ET_c for spring maize compared to the three independent models. The Nash–Sutcliffe efficiency coefficient (NSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were recorded as 0.956–0.958, 0.275–0.282 mm d⁻¹, 0.221–0.231 mm d⁻¹, and 0.957–0.961, respectively. The effectiveness of the hybrid RF-SVR-PSO model for daily ET_c estimation of spring maize in semi-arid regions has been substantiated. This aligns with research outcomes indicating that the mean bias error (MBE) under RF and Sc3 was 5.14 m³ ton⁻¹, the peak MBE under RF-SVR was 5.05 m³ ton⁻¹, and the minimal MBE under RF-SVR and Sc1 was 0.03. The highest Mean Absolute Relative Error (MARE) values for RF and XGB with Sc3 were 59.23 and 57.82, respectively, while the lowest values were recorded as −4.05 and −3.99 for SVR with Sc2 and Sc1 (Table 5).

5. Conclusions

There is a growing interest in improving agricultural water productivity due to the decline in water resources while meeting the increasing global demand for food using limited fresh water. The aim is to increase crop yields using less water, thus reducing the water footprint per unit of agricultural output. Accordingly, the research demonstrated the effectiveness of predicting the green water footprint of sugarcane crops in White Nile State, Sudan, to reach the best sustainable water management approach for the crop during the period 2001–2020 through the use of machine learning algorithms and remote sensing indices.

No significant advantage was observed when using hybrid models compared to single models, especially when dealing with the estimation of coefficient of determination (R^2). However, there was a significant difference in terms of RMSE, NBE, and MAE as performance indicators for GWFP prediction effectiveness, where the hybrid models had less errors compared to individual models.

The evaluation indices showed that Sc3 is the worst-case scenario, associated with remote sensing indices, while Sc1 (all parameters) and Sc4 ($P_{e_{eff}}$, T_{max} , T_{min} , and SA) were identified as the best-case scenarios, with convergence of the evaluation results. The choice between them depends on the availability of different criteria such as climatic parameters, crop parameters, and remote sensing indices. In cases of a lack of capabilities and limited data, the study recommends choosing Sc4.

The suboptimal functionality of Sc3 can be remedied by employing higher-resolution remote sensing datasets, increasing the temporal resolution, and integrating supplementary remote sensing indices, such as NDTI, GPP, and Soil Moisture along with LST to facilitate the understanding of the hydrological dynamics pertinent to sugarcane cultivation.

Author Contributions: Conceptualization, R.H.A.-T., M.E.A., A.A.-R.S.A., A.S. and A.M.; methodology, R.H.A.-T., M.E.A. and A.M.; software, R.H.A.-T., M.E.A. and A.M.; validation, R.H.A.-T., M.E.A., A.A.-R.S.A. and A.M.; formal analysis, R.H.A.-T., M.E.A., A.A.-R.S.A. and A.M.; investigation, A.A.-R.S.A., A.S., S.A.H., E.A.H. and A.M.; resources, A.A.-R.S.A., M.M.H., A.S., S.A.H. and E.A.H.; data curation, E.A.H. and A.S.; writing—original draft preparation, R.H.A.-T. and M.E.A.; writing—review and editing, R.H.A.-T., M.E.A., A.A.-R.S.A., M.M.H., A.S., S.A.H., E.A.H. and A.M.; project administration and funding acquisition, M.E.A., A.S. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Statement: All machine learning algorithms and model codes used in the research are sources available through the Internet for free, as well as satellite images of the study area obtained from Google Earth, and all data included in the research will be made available upon request, while the data from previous studies and research were obtained through the Cairo University platform, which provides research information on a regular basis.

Acknowledgments: The authors would like to express their thanks to the Faculty of Agriculture, Cairo University, and Faculty of Higher African Studies for their support in this work.

Conflicts of Interest: The authors declare no conflicts of interest. The authors declare they have no financial interests.

List of Abbreviations

Support vector regression (SVR); random forest (RF); extreme gradient boost (XGB) and artificial neural network (ANN); blue water footprint (BWF); Nash–Sutcliffe model efficiency coefficient (NSE); root mean square error (RMSE); the mean absolute error (MAE); the mean bias error (MBE); the coefficient of determination (R^2); mean average percentage error (MAPE); uncertainty with 95% (U_{95}).

References

1. Mohamed, E.S.E. Climate Change, Agricultural Production and Food Security in Sudan. *J. Econ. Res.* **2022**, *3*, 1–19. [\[CrossRef\]](#)
2. Verma, K.K.; Song, X.-P.; Yadav, G.; Degu, H.D.; Parvaiz, A.; Singh, M.; Huang, H.-R.; Mustafa, G.; Xu, L.; Li, Y.-R. Impact of agroclimatic variables on proteogenomics in sugar cane (*Saccharum* spp.) plant productivity. *ACS Omega* **2022**, *7*, 22997–23008. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Hassan Dahab, M.; Kheiry, A.N.; Abdalla, O.A. Energy Use Efficiency of Sugar Cane Production in the Central Clay Plain of Kenana Area. *J. Energy Res. Rev.* **2022**, *10*, 18–25. [\[CrossRef\]](#)
4. Hoekstra, A.Y. *The Water Footprint Assessment Manual: Setting the Global Standard*; Routledge: London, UK, 2011.
5. Hoekstra, A.Y. *The Water Footprint of Modern Consumer Society*; Routledge: London, UK, 2019.
6. El-Marsafawy, S.M.; Mohamed, A.I. Water footprint of Egyptian crops and its economics. *Alex. Eng. J.* **2021**, *60*, 4711–4721. [\[CrossRef\]](#)
7. Morillo, J.G.; Díaz, J.A.R.; Camacho, E.; Montesinos, P. Linking water footprint accounting with irrigation management in high value crops. *J. Clean. Prod.* **2015**, *87*, 594–602. [\[CrossRef\]](#)
8. Mekonnen, M.M.; Gerbens-Leenes, W. The water footprint of global food production. *Water* **2020**, *12*, 2696. [\[CrossRef\]](#)
9. Xu, Z.; Chen, X.; Wu, S.R.; Gong, M.; Du, Y.; Wang, J.; Li, Y.; Liu, J. Spatial-temporal assessment of water footprint, water scarcity and crop water productivity in a major crop production region. *J. Clean. Prod.* **2019**, *224*, 375–383. [\[CrossRef\]](#)

10. Cao, X.; Zeng, W.; Wu, M.; Li, T.; Chen, S.; Wang, W. Water resources efficiency assessment in crop production from the perspective of water footprint. *J. Clean. Prod.* **2021**, *309*, 127371. [CrossRef]
11. Quinteiro, P.; Ridoutt, B.G.; Arroja, L.; Dias, A.C. Identification of methodological challenges remaining in the assessment of a water scarcity footprint: A review. *Int. J. Life Cycle Assess.* **2018**, *23*, 164–180. [CrossRef]
12. Naranjo-Merino, C.A.; Ortíz-Rodríguez, O.O.; Villamizar-G, R.A. Assessing green and blue water footprints in the supply chain of cocoa production: A case study in the northeast of Colombia. *Sustainability* **2017**, *10*, 38. [CrossRef]
13. Abdel-Hameed, A.M.; Abuarab, M.E.-S.; Al-Ansari, N.; Mehawed, H.S.; Kassem, M.A.; He, H.; Gyasi-Agyei, Y.; Mokhtar, A. Winter potato water footprint response to climate change in Egypt. *Atmosphere* **2022**, *13*, 1052. [CrossRef]
14. Ray, A.S. Remote sensing in agriculture. *Int. J. Environ. Agric. Biotechnol.* **2016**, *1*, 238540. [CrossRef]
15. Khanal, S.; Kc, K.; Fulton, J.P.; Shearer, S.; Ozkan, E. Remote sensing in agriculture—Accomplishments, limitations, and opportunities. *Remote Sens.* **2020**, *12*, 3783. [CrossRef]
16. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. [CrossRef]
17. Saha, D.; Annamalai, M. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review Current Research in Food Science. *IEEE J.* **2021**, *16*, 4566–4578.
18. Antonopoulos, A.S.; Boutsikou, M.; Simantiris, S.; Angelopoulos, A.; Lazaros, G.; Panagiotopoulos, I.; Oikonomou, E.; Kanoupaki, M.; Tousoulis, D.; Mohiaddin, R.H. Machine learning of native T1 mapping radiomics for classification of hypertrophic cardiomyopathy phenotypes. *Sci. Rep.* **2021**, *11*, 23596. [CrossRef]
19. Zhang, K.; Li, Y.; Yu, Z.; Yang, T.; Xu, J.; Chao, L.; Ni, J.; Wang, L.; Gao, Y.; Hu, Y. Xin'anjiang nested experimental watershed (XAJ-NEW) for understanding multiscale water cycle: Scientific objectives and experimental design. *Engineering* **2022**, *18*, 207–217. [CrossRef]
20. Mirani, A.; Memon, M.S.; Chohan, R.; Wagan, A.A.; Qabulio, M. Machine learning in agriculture: A review. *Int. J. Sci. Technol. Res.* **2021**, *10*, 5.
21. Kumar, V.; Yadav, S. Optimization of cropping patterns using elitist-Jaya and elitist-TLBO algorithms. *Water Resour. Manag.* **2019**, *33*, 1817–1833. [CrossRef]
22. Sun, J.; Lai, Z.; Di, L.; Sun, Z.; Tao, J.; Shen, Y. Multilevel deep learning network for county-level corn yield estimation in the us corn belt. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5048–5060. [CrossRef]
23. Cho, S.; Vasarhelyi, M.A.; Sun, T.; Zhang, C. *Learning from Machine Learning in Accounting and Assurance*; American Accounting Association: Lakewood Ranch, FL, USA, 2020; pp. 1–10.
24. Veeragandham, S.; Santhi, H. A review on the role of machine learning in agriculture. *Scalable Comput. Pract. Exp.* **2020**, *21*, 583–589. [CrossRef]
25. Everingham, Y.; Sexton, J.; Skocaj, D.; Inman-Bamber, G. Accurate prediction of sugarcane yield using a random forest algorithm. *Agron. Sustain. Dev.* **2016**, *36*, 1–9. [CrossRef]
26. Okal, H.A.; Ngetich, F.K.; Okeyo, J.M. Spatio-temporal characterisation of droughts using selected indices in Upper Tana River watershed, Kenya. *Sci. Afr.* **2020**, *7*, e00275. [CrossRef]
27. Mompremier, R.; Her, Y.; Hoogenboom, G.; Migliaccio, K.; Muñoz-Carpena, R.; Brym, Z.; Colbert, R.; Jeune, W. Modeling the response of dry bean yield to irrigation water availability controlled by watershed hydrology. *Agric. Water Manag.* **2021**, *243*, 106429. [CrossRef]
28. Adhikari, N.D.; Simko, I.; Mou, B. Phenomic and physiological analysis of salinity effects on lettuce. *Sensors* **2019**, *19*, 4814. [CrossRef]
29. Yao, J.; Wu, J.; Xiao, C.; Zhang, Z.; Li, J. The classification method study of crops remote sensing with deep learning, machine learning, and Google Earth engine. *Remote Sens.* **2022**, *14*, 2758. [CrossRef]
30. Lillesand, T.M.; Kiefer, R.W.; Chipman, J.W. *Remote Sensing and Image Interpretation*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2004.
31. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
32. Rouse, J.H.; Shaw, J.A.; Lawrence, R.L.; Lewicki, J.L.; Dobeck, L.M.; Repasky, K.S.; Spangler, L.H. Multi-spectral imaging of vegetation for detecting CO₂ leaking from underground. *Environ. Earth Sci.* **2010**, *60*, 313–323. [CrossRef]
33. Hub, S. NDVI (Normalized Difference Vegetation Index). 2018. Available online: <https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndvi/> (accessed on 1 January 2023).
34. Mokhtar, A.; He, H.; Alsafadi, K.; Li, Y.; Zhao, H.; Keo, S.; Bai, C.; Abuarab, M.; Zhang, C.; Elbagoury, K.; et al. Evapotranspiration as a response to climate variability and ecosystem changes in southwest, China. *Environ. Earth Sci.* **2020**, *79*, 312. [CrossRef]
35. Mokhtar, A.; He, H.; Nabil, M.; Kouadri, S.; Salem, A.; Elbeltagi, A. Securing China's rice harvest: Unveiling dominant factors in production using multi-source data and hybrid machine learning models. *Sci. Rep.* **2024**, *14*, 14699. [CrossRef]
36. Mokhtar, A.; Hamed, M.M.; He, H.; Salem, A.; Hendy, Z.M. Egypt's water future: AI predicts evapotranspiration shifts across climate zones. *J. Hydrol. Reg. Stud.* **2024**, *56*, 101968. [CrossRef]
37. Allen, R.G. Using the FAO-56 dual crop coefficient method over an irrigated region as part of an evapotranspiration intercomparison study. *J. Hydrol.* **2000**, *229*, 27–41. [CrossRef]

38. Sumner, D.M.; Jacobs, J.M. Utility of Penman–Monteith, Priestley–Taylor, reference evapotranspiration, and pan evaporation methods to estimate pasture evapotranspiration. *J. Hydrol.* **2005**, *308*, 81–104. [[CrossRef](#)]
39. Chico, D.; Aldaya, M.M.; Garrido, A. A water footprint assessment of a pair of jeans: The influence of agricultural policies on the sustainability of consumer products. *J. Clean. Prod.* **2013**, *57*, 238–248. [[CrossRef](#)]
40. Hoekstra, A.Y.; Chapagain, A.; Martinez-Aldaya, M.; Mekonnen, M. *Water Footprint Manual: State of the Art 2009*; Water Footprint Network: Enschede, The Netherlands, 2009.
41. Li, Z.; Xu, X.; Yu, B.; Xu, C.; Liu, M.; Wang, K. Quantifying the impacts of climate and human activities on water and sediment discharge in a karst region of southwest China. *J. Hydrol.* **2016**, *542*, 836–849. [[CrossRef](#)]
42. Djebou, D.C.S.; Singh, V.P. Impact of climate change on precipitation patterns: A comparative approach. *Int. J. Climatol.* **2016**, *36*, 3588–3606. [[CrossRef](#)]
43. Li, R.; Xiong, L.; Xiong, B.; Li, Y.; Xu, Q.; Cheng, L.; Xu, C.-Y. Investigating the downstream sediment load change by an index coupling effective rainfall information with reservoir sediment trapping capacity. *J. Hydrol.* **2020**, *590*, 125200. [[CrossRef](#)]
44. Mekonnen, M.M.; Hoekstra, A.Y. The green, blue and grey water footprint of crops and derived crop products. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1577–1600. [[CrossRef](#)]
45. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
46. Chen, T.; Zhu, L.; Niu, R.-q.; Trinder, C.J.; Peng, L.; Lei, T. Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *J. Mt. Sci.* **2020**, *17*, 670–685. [[CrossRef](#)]
47. Chutia, D.; Borah, N.; Baruah, D.; Bhattacharyya, D.K.; Raju, P.; Sarma, K. An effective approach for improving the accuracy of a random forest classifier in the classification of Hyperion data. *Appl. Geomat.* **2020**, *12*, 95–105. [[CrossRef](#)]
48. Ghorbani, M.A.; Deo, R.C.; Kim, S.; Hasanpour Kashani, M.; Karimi, V.; Izadkhan, M. Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia. *Soft Comput.* **2020**, *24*, 12079–12090. [[CrossRef](#)]
49. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
50. Ferreira, L.B.; da Cunha, F.F. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. *Comput. Electron. Agric.* **2020**, *178*, 105728. [[CrossRef](#)]
51. Zheng, H.; Yuan, J.; Chen, L. Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. *Energies* **2017**, *10*, 1168. [[CrossRef](#)]
52. Suo, G.; Song, L.; Dou, Y.; Cui, Z. Multi-dimensional short-term load forecasting based on XGBoost and fireworks algorithm. In Proceedings of the 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Wuhan, China, 8–10 November 2019; pp. 245–248.
53. Shamsudin, H.; Sabudin, M.; Yusof, U.K. Hybridisation of RF (Xgb) to improve the tree-based algorithms in learning style prediction. *IAES Int. J. Artif. Intell.* **2019**, *8*, 422. [[CrossRef](#)]
54. Hou, W.; Yin, G.; Gu, J.; Ma, N. Estimation of spring maize evapotranspiration in semi-arid regions of Northeast China using machine learning: An improved SVR Model based on PSO and RF algorithms. *Water* **2023**, *15*, 1503. [[CrossRef](#)]
55. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [[CrossRef](#)]
56. Springmann, M.; Mason-D’Croz, D.; Robinson, S.; Wiebe, K.; Godfray, H.C.J.; Rayner, M.; Scarborough, P. Health-motivated taxes on red and processed meat: A modelling study on optimal tax levels and associated health impacts. *PLoS ONE* **2018**, *13*, e0204139. [[CrossRef](#)]
57. Gueymard, C.A. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. *Renew. Sustain. Energy Rev.* **2014**, *39*, 1024–1034. [[CrossRef](#)]
58. Behar, O.; Khellaf, A.; Mohammedi, K. A novel parabolic trough solar collector model—Validation with experimental data and comparison to Engineering Equation Solver (EES). *Energy Convers. Manag.* **2015**, *106*, 268–281. [[CrossRef](#)]
59. Li, D.; Liu, Y.; Zhang, X. Linear statics and free vibration sensitivity analysis of the composite sandwich plates based on a layerwise/solid-element method. *Compos. Struct.* **2013**, *106*, 175–200. [[CrossRef](#)]
60. Downing, A.R.; Greenberg, I.B.; Peha, J.M. OSCAR: A system for weak-consistency replication. In Proceedings of the 1990 Proceedings. Workshop on the Management of Replicated Data, Houston, TX, USA, 8–9 November 1990; pp. 26–30.
61. Balaghi, R.; Tychon, B.; Eerens, H.; Jlibene, M. Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. *Int. J. Appl. Earth Obs. Geoinf.* **2008**, *10*, 438–452. [[CrossRef](#)]
62. Tao, M.; Zhang, T.; Xie, X.; Liang, X. Water footprint modeling and forecasting of cassava based on different artificial intelligence algorithms in Guangxi, China. *J. Clean. Prod.* **2023**, *382*, 135238. [[CrossRef](#)]
63. Mokhtar, A.; El-Ssawy, W.; He, H.; Al-Anasari, N.; Sammen, S.S.; Gyasi-Agyei, Y.; Abuarab, M. Using machine learning models to predict hydroponically grown lettuce yield. *Front. Plant Sci.* **2022**, *13*, 706042. [[CrossRef](#)]
64. Abdel-Hameed, A.M.; Abuarab, M.; Al-Ansari, N.; Sayed, H.; Kassem, M.A.; Elbeltagi, A.; Mokhtar, A. Estimation of Potato Water Footprint Using Machine Learning Algorithm Models in Arid Regions. *Potato Res.* **2024**, *67*, 1755–1774. [[CrossRef](#)]
65. Ge, J.; Zhao, L.; Yu, Z.; Liu, H.; Zhang, L.; Gong, X.; Sun, H. Prediction of greenhouse tomato crop evapotranspiration using XGBoost machine learning model. *Plants* **2022**, *11*, 1923. [[CrossRef](#)]

66. Elbeltagi, A.; Deng, J.; Wang, K.; Hong, Y. Crop Water footprint estimation and modeling using an artificial neural network approach in the Nile Delta, Egypt. *Agric. Water Manag.* **2020**, *235*, 106080. [[CrossRef](#)]
67. Wu, L.; Huang, G.; Fan, J.; Ma, X.; Zhou, H.; Zeng, W. Hybrid extreme learning machine with meta-heuristic algorithms for monthly pan evaporation prediction. *Comput. Electron. Agric.* **2020**, *168*, 105115. [[CrossRef](#)]
68. Azzam, A.; Zhang, W.; Akhtar, F.; Shaheen, Z.; Elbeltagi, A. Estimation of green and blue water evapotranspiration using machine learning algorithms with limited meteorological data: A case study in Amu Darya River Basin, Central Asia. *Comput. Electron. Agric.* **2022**, *202*, 107403. [[CrossRef](#)]
69. Elhussiny, K.T.; Hassan, A.M.; Habssa, A.A.; Mokhtar, A. Prediction of water distribution uniformity of sprinkler irrigation system based on machine learning algorithms. *Sci. Rep.* **2023**, *13*, 20885. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.