agronomy

MDPI

Article

Inversion of Soil Moisture Content in Cotton Fields Using GBR-RF Algorithm Combined with Sentinel-2 Satellite Spectral Data

Xu Li [1,†], Jingming Wu [1,†], Jun Yu [2,*], Zhengli Zhou [2], Qi Wang [1], Wenbo Zhao [1] and Lijun Hu [1]

[1] Key Laboratory of Tarim Oasis Agriculture, Ministry of Education, College of Information Engineering, Tarim University, Alar 843300, China; lixu2866830@gmail.com (X.L.); 10757222273@stumail.taru.edu.cn (J.W.); qiw115986@gmail.com (Q.W.); suifengdefeng0505@gmail.com (W.Z.); lijunhu740@gmail.com (L.H.)
[2] Faculty of Horticulture and Forestry, Tarim University, Alar 843300, China; zhenglizhou136@gmail.com
* Correspondence: yuj860168@gmail.com
† These authors contributed equally to this work.

Abstract: Soil moisture content plays a vital role in agricultural production, significantly influencing crop growth, development, and yield. Thoroughly understanding the specific soil moisture content in cotton fields is crucial for enhancing agricultural efficiency and driving sustainable agricultural development. This study utilized the gradient-boosting regression–random forest (GBR-RF) algorithm and the GBR and RF algorithms separately, in conjunction with Sentinel-2 satellite images, to estimate cotton soil moisture content, focusing on the B1–B8 bands and in particular the sensitive B6, B7, and B8 bands. The soil data in the jujube orchard of the study area were collected using soil augers at a depth of 30 cm, with soil data collected from a depth of 20 to 30 cm. The findings revealed that the integrated learning algorithm GBR-RF demonstrated high accuracy, with R^2, MAE, and MSE results of 0.8838, 1.0121, and 1.6168, respectively. In comparison, the results using just the GBR algorithm yielded R^2, MAE, and MSE values of 0.8158, 1.1327, and 1.9645, respectively, while those obtained from the RF algorithm were 0.8415, 1.0680, and 1.8331, respectively. These results indicate that the algorithms exhibited strong generalization, robustness, and accuracy, with GBR-RF outperforming GBR and RF by 8.34% and 5.03%, respectively, in combination with using the B1–B8 bands for inversion. Furthermore, utilizing the full-band data resulted in R^2 values that were up to 24.27% higher than those of the individual bands, affirming the efficacy of band combinations for improved accuracy. This study's demonstration of the positive impact of integrated learning algorithms on estimating cotton soil moisture content underscores the advantages of multi-band data combinations over single-band data, highlighting their ability to enhance accuracy without significantly impacting errors. Importantly, this study's findings, while not limited to a single experimental field, have broad applicability in cotton precision agriculture, offering valuable insights for research on yield enhancement and agricultural efficiency.

Keywords: cotton; soil moisture content; Sentinel-2; spectral data; gradient-boosting regression (GBR); random forest (RF)

Citation: Li, X.; Wu, J.; Yu, J.; Zhou, Z.; Wang, Q.; Zhao, W.; Hu, L. Inversion of Soil Moisture Content in Cotton Fields Using GBR-RF Algorithm Combined with Sentinel-2 Satellite Spectral Data. *Agronomy* **2024**, *14*, 784. https://doi.org/10.3390/agronomy14040784

Academic Editor: Belen Gallego-Elvira

Received: 29 February 2024
Revised: 6 April 2024
Accepted: 8 April 2024
Published: 10 April 2024

1. Introduction

Xinjiang is recognized as one of the most important cotton-producing areas in China and serves as a significant global cotton production base. Among the various cash crops cultivated in Xinjiang, cotton holds a particularly prominent position due to its economic significance. The development of the cotton industry in Xinjiang has played a vital role in fostering local economic growth, boosting farmers' income, and generating employment opportunities [1].

Cotton (scientific name: Gossypium hirsutum) is an important cash crop belonging to the genus Cotton in the family Mallowaceae [2]. The fruit of cotton is known as a boll

and contains cotton seeds, which can be processed to extract cotton fiber, one of the main raw materials for manufacturing textiles. Cotton growth is adaptable to a wide range of climatic and soil requirements, but tropical and subtropical climates are preferred. The typical growing period is 4 to 6 months, requiring sufficient sunlight and water during growth [3]. It is a perennial herbaceous plant but is generally cultivated as an annual plant. Cotton, as one of the major fiber crops in the world, is versatile and important [4].

In agricultural production, soil moisture is a critical factor that directly influences crop growth and yield [5]. Throughout the growing season, cotton is a water-consuming crop and necessitates large amounts of water to support normal growth and development [6]. Maintaining adequate soil moisture is crucial, as it facilitates seed germination, seedling growth, vegetative development, and the establishment of a healthy root system, ultimately enhancing both the yield and quality of the cotton. Optimal soil moisture levels play a vital role in regulating the physiological processes of plants. An imbalance in soil moisture—whether due to deficiency or excess—can disrupt these physiological processes, leading to adverse effects on the growth and yield of cotton [7]. Therefore, the precise monitoring and management of soil moisture content during the growth stages of cotton are paramount for achieving high-yield, quality, and consistent cotton production. The strategic utilization of soil moisture resources can effectively regulate the growth environment of cotton, leading to improved yield and quality and subsequently maximizing economic returns [8].

Various remote sensing techniques, including satellite remote sensing, have been extensively employed for soil moisture monitoring and estimation in recent decades. Satellite remote sensing has garnered significant interest for its ability to provide wide coverage, frequent periodic observations, and non-contact acquisition of information. Despite these advantages, accurately obtaining soil moisture information remains a challenge [9,10]. The importance of using satellite images to map soil moisture lies in providing comprehensive and timely soil moisture monitoring data for cotton growers, enhancing their decision-making processes, and optimizing cultivation management [11]. Soil moisture also has a direct impact on the occurrence of some diseases and pests. Monitoring soil moisture allows for the timely detection of overly wet or dry soil conditions, enabling appropriate measures to adjust soil moisture levels promptly, thus reducing the incidence of diseases and pests, as well as ensuring the healthy growth of cotton [12]. Soil moisture mapping can also help growers better understand the distribution of nutrients in the soil. Based on soil moisture information, growers can adjust their fertilization plan to ensure that cotton receives a sufficient nutrient supply during growth, thereby enhancing yield and quality [13]. In conclusion, soil moisture mapping plays a crucial role in cotton cultivation. It not only assists growers in optimizing water management, irrigation, fertilization, and other agricultural activities to increase cotton yield and quality but also mitigates the occurrence of diseases and pests, reduces disaster risks, and promotes the sustainable development of the cotton industry.

In recent years, research on the inversion of soil moisture and other related indicators based on machine learning algorithms has drawn significant attention [14,15]. Remote sensing methods provide high-resolution spectral data, offering a wealth of information for the inversion of soil moisture content [16,17]. Kingsley John et al. successfully estimated the variability in soil organic carbon in alluvial soils using machine learning algorithms in conjunction with environmental variables and soil nutrient indicators, achieving an optimal R^2 value of 0.68 [18]. Felipe B. de Santana et al. conducted a study comparing the effectiveness of partial least squares (PLSs) and support vector machine (SVM) models in predicting soil organic matter and particle size using a visible–near-infrared spectral library, resulting in a remarkable 22% reduction in the RMSEP value [19]. Xiaoping Wang et al. employed a fractional-order filtering algorithm and grid-searching support vector machine modeling to extract soil salinity information, achieving an accuracy rate of 91.9% [20]. Coleen Carranza et al. utilized random forest to estimate soil moisture in the root zone and found that random forest outperformed process-based models [21]. Yue Zhang et al.

employed a random forest model with different predictors to compare the mapping of total soil nitrogen storage from remote sensing data, concluding that the random forest method accurately captures alterations in soil total nitrogen content [22]. Moreover, Mohammad Hosseinpour-Zarnaq et al. employed vis–NIR spectral data along with a convolutional neural network (CNN) model to predict soil properties, demonstrating the feasibility of using near-infrared spectral data as a swift and non-invasive tool for assessing soil properties [23]. Lei Zhang et al. developed a CNN–long short-term memory (CNN-LSTM) model to predict soil organic carbon content based on a long-term series of climatic variables from MODIS, and their research indicated that remote sensing methods utilizing random forest accurately capture changes in soil total nitrogen content, showcasing the promising potential of hybrid deep learning models [24]. Despite the recent increase in related research, a comprehensive method regarding the relevant algorithms has yet to be summarized, and issues such as spectral correlation features continue to be subject to ongoing investigation.

This study aims to establish an accurate and reliable model for the real-time monitoring and estimation of soil moisture content during the cotton growth period by utilizing the GBR-RF algorithm in combination with Sentinel-2 satellite spectral data. The feasibility and advantages of this approach are explored through an analysis of spectral data from different bands and their integration with ground-based measurements of soil moisture content. The results of this study are expected to offer valuable reference information for agricultural production management. This information will aid in optimizing agricultural production decisions, enhancing crop yield and quality, and promoting the sustainable use and protection of land resources.

This paper begins by providing an overview of the importance of soil moisture content inversion. Following this, the characteristics and acquisition of the GBR-RF algorithm and Sentinel-2 satellite spectral data are briefly discussed. Subsequently, the research methodology and experimental design employed in this study are described. Finally, the anticipated outcomes of this study and its potential applications in agricultural production are considered.

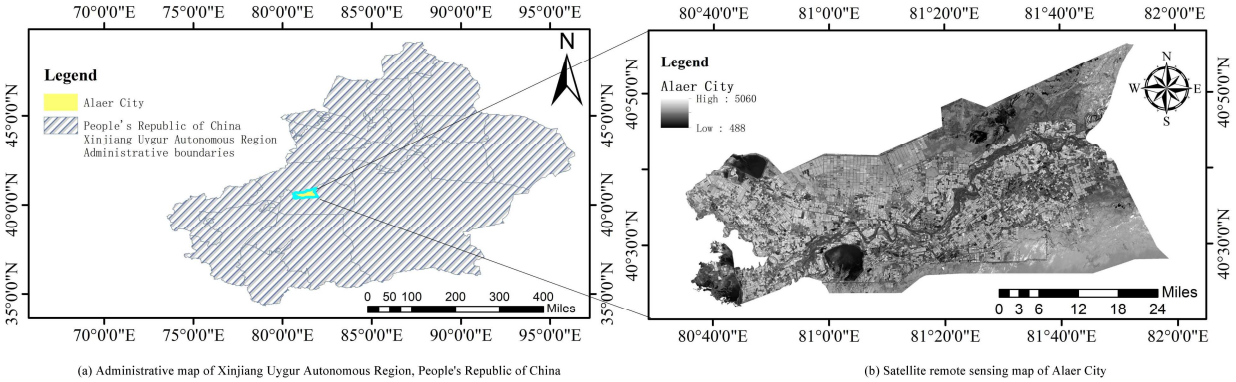
2. Materials and Methods

2.1. Overview and Data Collection in the Experimental Zone

Located in Alar City, Xinjiang Uygur Autonomous Region (XUAR), the research pilot area is subject to the joint administration of the XUAR and the Xinjiang Production and Construction Corps (XPC), and it follows a division–municipality management system. Encompassing an expanse of 6923.4 square kilometers, it experiences a temperate, extreme continental, and arid desert climate. Geographically positioned in the southern foothills of the Tianshan Mountains, at the northern boundary of the Taklamakan Desert and the upper reaches of the Tarim River, where it converges with the Aksu River, Hotan River, and Yarkant River, the region is located between longitudes 80°30′ and 81°58′ and latitudes 40°22′ and 40°57′. The terrain is characterized by the alluvial fine-soil plains of the Tarim River, featuring a gentle uplift along the riverbanks and the sides of the alluvial gullies, with a topography sloping from northwest to southeast. Notably, the area experiences an extreme minimum temperature of −33.2 °C, with an average annual solar radiation range of 133.7–146.3 kcal/cm^2. Furthermore, the average annual sunshine duration varies from 2556.3 to 2991.8 h, while precipitation is scarce, with minimal winter snowfall. The region faces intense surface evaporation, with an average annual precipitation rate ranging from 40.1 to 82.5 mm and an average annual evaporation rate between 1876.6 and 2558.9 mm.

The conducted experiment involved collecting soil samples from cotton fields and determining their water content in addition to analyzing the relevant spectral data from the Sentinel-2 satellite for inversion. Satellite remote sensing image data were acquired from the Google Earth Engine (GEE) platform, with the satellite of choice being Sentinel-2 at Level-2A, captured between 1 April and 10 April, and, due to the correlation between soil data and bands, B1–B8 band data were also acquired. European Space Agency satellite

data were used in combination with ENVI for data processing; a comparative analysis on the GEE platform indicated its relative effectiveness, leading to its selection for the study. To enhance the data quality, cloud removal, atmospheric correction, and radiometric calibration were performed within the platform before proceeding with the subsequent research using remote sensing images. Soil samples were collected from cotton fields in the vicinity of Aral City cultivated by local farmers. The data used in this experiment were collected on 7 April 2023, using soil augers to collect soil samples at depths ranging from 0 cm to 30 cm. The study focused on the 20-cm to 30-cm soil data for experimentation, with a total of 180 soil samples collected. The soil samples were sealed in plastic bags for storage, and their wet weights were measured using a digital balance accurate to the nearest one hundredth of a gram. Subsequently, the samples were dried in an oven to obtain their dry weights, and the moisture content was calculated using the appropriate formula. The equation for soil water content is shown in Equation (1). θ denotes soil moisture content (%), m_w denotes the mass of wet soil (g), and m_d denotes the mass of dry soil (g). The study area map is depicted in Figure 1.



(a) Administrative map of Xinjiang Uygur Autonomous Region, People's Republic of China

(b) Satellite remote sensing map of Alaer City

Figure 1. Study area map—Alar City, Xinjiang Uygur Autonomous Region, People's Republic of China.

$$\theta = \frac{m_w - m_d}{m_d} \times 100\% \quad (1)$$

Google Earth Engine (GEE) is a cloud platform developed by Google for geoscience data analysis and processing. It provides a large number of geoscience datasets, powerful computational resources, and easy-to-use programming interfaces that enable users to perform large-scale geoscience data analysis and processing in the cloud [25].

2.2. Data Processing

In this experiment, the band data of the satellite remote sensing images were acquired and output using ENVI. It is important to highlight that the remote sensing images were acquired at a 10-m resolution. In this experiment, ground samples were collected at 10-m intervals, aligning each sample point with a pixel of the same resolution. Although the data acquisition interval selected for this study matches the resolution, it is essential to recognize that there may exist a certain level of error, as the data may not precisely correspond to each individual pixel point within the image. Nonetheless, the data collection points in this study are situated in the experimental field in close proximity to its center, with distances exceeding 20 m between the data collection points and the field's edges. Consequently, the experimental data obtained were deemed valid within the scope of this study. Furthermore, the crops grown in the sampled locations in this study were all one-year-old cotton plants in the seedling stage. Since they were cultivated in open fields within the same county-level city, the growth conditions across the research areas were essentially the same.

Due to the data utilized in this experiment being directly extracted band data, the output is a single value rather than a continuous whole group of data. Consequently, the attempted preprocessing of the data using techniques such as first-order derivatives

and inverses yielded manifestly unreasonable results. Consequently, the outcomes of the attempted inversions were notably poor. Therefore, the operations intended to manipulate the data were not pursued further in this study.

The data were cleaned both manually and using machine learning algorithms, following data cleaning instructions. No detectable anomalies were found, so further explanation of these operations is omitted.

2.3. Pearson Correlation Analysis

A Pearson correlation analysis is a statistical method utilized to evaluate the linear correlation between two continuous variables. It calculates the Pearson correlation coefficient to determine the strength and direction of the correlation between the two variables. The coefficient ranges from -1 to 1, with values closer to these extremes indicating a stronger correlation, while those closer to 0 indicate a weaker correlation. In this experiment, manipulation was carried out to understand the correlations between the variables in order to aid in modeling and determine which variables were important for explaining the changes in the target variable. The main objective was to identify the key factors affecting the results, given the large number of overall independent variables. During the correlation analysis in this study, the multispectral data were used as the independent variable data and the soil moisture content as the dependent variable data.

2.4. Gradient-Boosting Regression (GBR)

The gradient boosting machine method, initially proposed by Jerome Friedman, led to the development of gradient-boosting regression [26]. Building upon this method, Trevor Hastie, Robert Tibshirani, and Jerome Friedman introduced the gradient boosting tree method, which integrates gradient boosting with a decision tree model to enhance prediction accuracy. Gradient-boosting regression involves iteratively training a series of weak prediction models, each of which aims to correct the residuals of the preceding model. With each iteration, the model fits the residuals, compares predicted values against observed data, and determines the next optimization direction. Through this progressive process, the model gradually refines the prediction accuracy of the target variable.

For this experiment, we chose the gradient-boosting regression algorithm due to its several advantages. This algorithm is known for its high prediction accuracy, insensitivity to outliers, and flexibility. Importantly, the issues related to long training times and cumbersome parameter tuning are successfully addressed by constraining specific parameters, resulting in improved efficiency without compromising prediction accuracy.

2.5. Random Forest (RF)

The history of the random forest algorithm can be traced back to the original idea proposed by Leo Breiman. Random forest (RF) is an integrated learning method based on decision tree construction, and it was further developed and shaped by Breiman et al. in their influential 2004 paper "Random Forests" [27]. Random forests enhance prediction accuracy and stability by training multiple decision trees and then integrating their outputs.

In this study, the random forest algorithm is utilized due to its high accuracy and robustness resulting from its good predictive performance, efficient data processing capabilities, and unique ability to evaluate the importance of features. Random forest is particularly advantageous because it can effectively handle missing values without requiring data normalization. This underscores the algorithm's versatility and reliability in analyzing complex datasets.

2.6. GBR-RF

In this experiment, an integrated learning approach combining the GBR and RF algorithms is used to address their individual limitations and leverage their strengths effectively.

Both RF and GBR are high-performance algorithms that excel in handling various data types and complex feature relationships. These models can effectively tackle both

numerical and categorical data while remaining insensitive to feature scaling, making them ideal for heterogeneous datasets. By combining these two models, their individual strengths can be leveraged to enhance the model's generalization ability, thereby improving its robustness in predicting new data. The randomness inherent in RF, coupled with the integrated learning concept, helps mitigate the risk of overfitting, while gradient-boosted regression's stepwise fitting of residuals aids in controlling the model's complexity to prevent overfitting on training data. Furthermore, the combined use of these algorithms offers a more comprehensive assessment of feature importance, allowing for a deeper understanding of the impact of individual features on prediction results. In essence, the synergy between RF and GBR not only enhances predictive performance but also provides valuable insights into the prediction process, making it a powerful tool for data analysis and modeling.

In this study, data were first read and processed using the pandas library. Next, the random forest algorithm was employed to evaluate and select important features to identify the most representative ones. Subsequently, a ridge regression algorithm was utilized to build a regression model, which was then trained and predicted based on the selected features. The algorithms were implemented using the scikit-learn library. Finally, model performance was assessed by computing metrics such as mean squared error, and visualizations were generated using the matplotlib library to intuitively depict the relationship between model predictions and actual labels, aiding in the analysis of the accuracy and applicability of the model.

In summary, the integrated approach of combining the GBR and RF algorithms offers several advantages that are crucial in addressing the research problem at hand. These advantages include a stronger generalization ability, better handling of heterogeneous data, a reduced risk of overfitting, high performance, and flexibility, along with the ability to conduct a feature importance analysis. This integrated approach has been demonstrated to yield significant success in practical applications, making it a suitable and effective solution for the research problem under examination.

2.7. Assessment of Indicators

In this experiment, we use R squared (R^2) [28], mean absolute error (MAE), and mean squared error (MSE) [29] as evaluation metrics to compare the data inversion results across different algorithms and wavelength bands.

R^2 is a common metric used to assess the goodness of fit of a regression model. It indicates the proportion of the variance in the dependent variable (target variable) that can be explained by the independent variable (characteristic variable). The specific formula for R^2 is shown in Equation (2). The value of R^2 ranges from 0 to 1, with values closer to 1 indicating that the model fits the data better and values closer to 0 indicating that the model fits the data worse. Here, $SS_{residual}$ denotes the sum of squares of the model residuals, and SS_{total} denotes the total sum of squares of the dependent variable.

MAE serves as a crucial metric for evaluating the prediction error of a regression model by calculating the average absolute deviation between the model's predicted value and the actual value. A smaller MAE value suggests a lower prediction error, making it a reliable measure of the model's predictive accuracy. Unlike MSE, MAE is less sensitive to outliers due to its utilization of absolute values, rendering it more appropriate for assessing models with outlier-ridden datasets. The specific formula for MAE, as represented in Equation (3), involves the true value, y_i, for the ith sample and the model's predicted value, $\overline{y}_i$, for that same sample, with "n" denoting the total number of samples.

MSE serves as a prevalent metric for evaluating the prediction error of a regression model by calculating the average squared deviation between the predicted values generated by the model and the actual true values. This metric plays a crucial role in quantifying the accuracy of the model's predictions, with a lower MSE value corresponding to a lower error in the predictive outcomes. Unlike mean absolute error (MAE), MSE exhibits greater sensitivity towards large errors as it computes the average of squared errors. The

specific mathematical expression representing MSE is denoted by Equation (4), with its interpretation aligning with that of Equation (3).

The reason for the simultaneous consideration of MAE and MSE is as follows: In this context, MAE primarily functions to assess the impact of outliers in the dataset. Despite the steps taken to address outliers in the data, as detailed in the preceding section, outliers were still observable in the box plots representing the experimental process. Consequently, MAE was integrated into the analysis workflow when the data re-entered the preprocessing stage, leading to consistent outcomes. Since MAE demonstrates a lower sensitivity to accuracy fluctuations than MSE, it was included for joint evaluation with the R^2 metric.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y}_i| \tag{3}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 \tag{4}$$

3. Results

3.1. Visualization of Experimental Data

In this experiment, data were collected from nine cotton plots at different locations, and 20 sets of soil were taken from each experimental plot. The corresponding coordinate data were recorded. Subsequently, the ENVI software was utilized to extract data from various bands of remote sensing images, which were then analyzed. The extracted information is visualized in Figure 2.

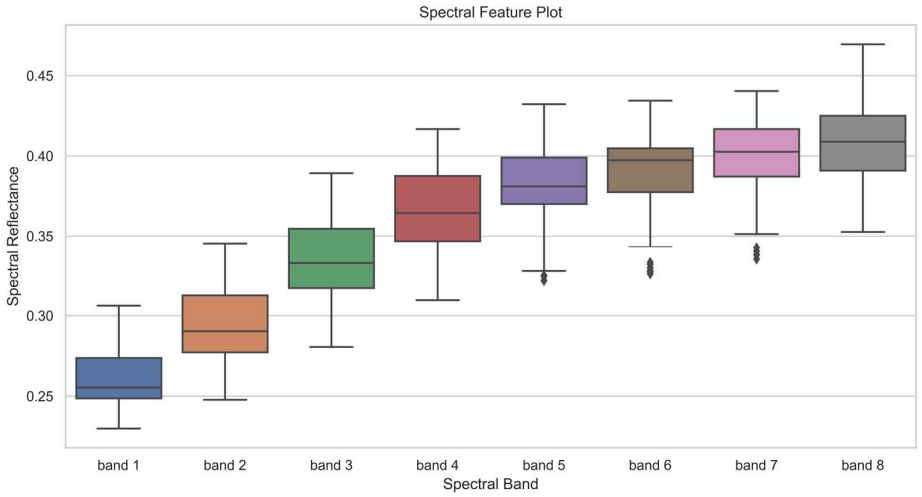


Figure 2. Box plots of satellite remote sensing multispectral band data for the study area.

3.2. Correlation Analysis

In order to conduct data inversion successfully, it is essential to combine all bands and analyze the correlation between the data from each band and soil moisture. This analysis helps in understanding how the variables influence each other and allows for adequate preparation for subsequent modeling and research endeavors. The correlation thermogram results, depicted in Figure 3, illustrate this relationship. Notably, bands B6, B7, and B8 exhibit higher sensitivity than the other bands. Accordingly, in addition to utilizing all bands for inversion, dedicated inversions were performed for these three bands individually, followed by a comprehensive analysis of the aggregated results.

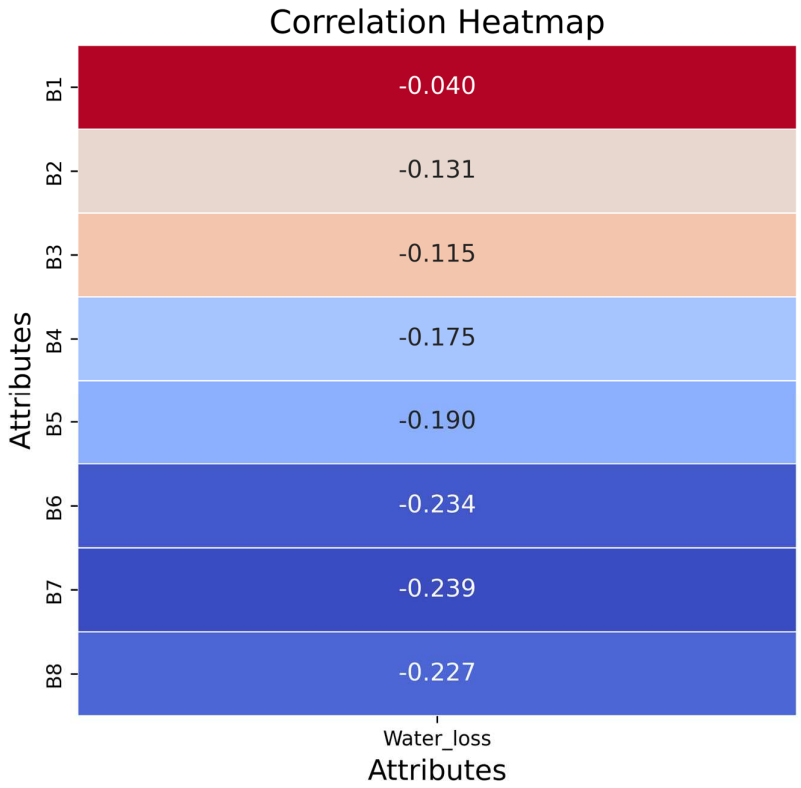


Figure 3. Correlation heatmap between satellite spectral band data and cotton soil moisture content.

3.3. Summary and Analysis of Soil Moisture Model Inversion Results for Cotton Land

In the following experiments, the data were preprocessed with specific bands and analyzed using the GBR, RF, and GBR-RF algorithms. The datasets were divided into a 70:30 ratio for training and validation. The results were then evaluated based on the analysis conducted.

In this study, the inversion process for cotton soil moisture content was carried out using a combination of various algorithms with the band data. Initially, the GBR algorithm was employed for experimentation, with the parameters adjusted to achieve optimal results. Specifically, a combination of 100 trees, a learning rate of 0.1, a maximum tree depth of 3, a minimum number of samples at internal nodes of 2, a least squares regression loss function, a minimum number of samples at leaf nodes of 1, and an alpha parameter of 0.9 were chosen. The final R^2 obtained was 0.8158, MAE was 1.1327, and MSE was 1.9645, with the detailed outcomes presented in Figure 4a. Subsequently, the RF algorithm was implemented with 100 trees, a tree depth of 5, a minimum number of samples at nodes of 2, and a minimum number of samples at leaf nodes of 1, resulting in an R^2 value of 0.8415, an MAE of 1.0680, and an MSE of 1.8331, as shown in Figure 4b. Furthermore, an integrated approach using the GBR-RF algorithms was adopted through a stacked regression model. This involved 100 trees and a random seed number of 42, with the final regressor being the RF model. The minimum number of samples at internal nodes for the RF component was set to 2, the learning rate for the GBR component was 0.1, and the maximum tree depth was limited to 3. The optimal feature importance threshold was established at 0.1, resulting in an R^2 value of 0.8838, an MAE of 1.0121, and an MSE of 1.6168, as illustrated in Figure 4c. Notably, the integrated learning algorithm of GBR-RF showed significantly higher performance than the individual GBR and RF algorithms in predicting cotton soil moisture content. Specifically, the RF algorithm consistently outperformed the GBR algorithm based on the correlation results.

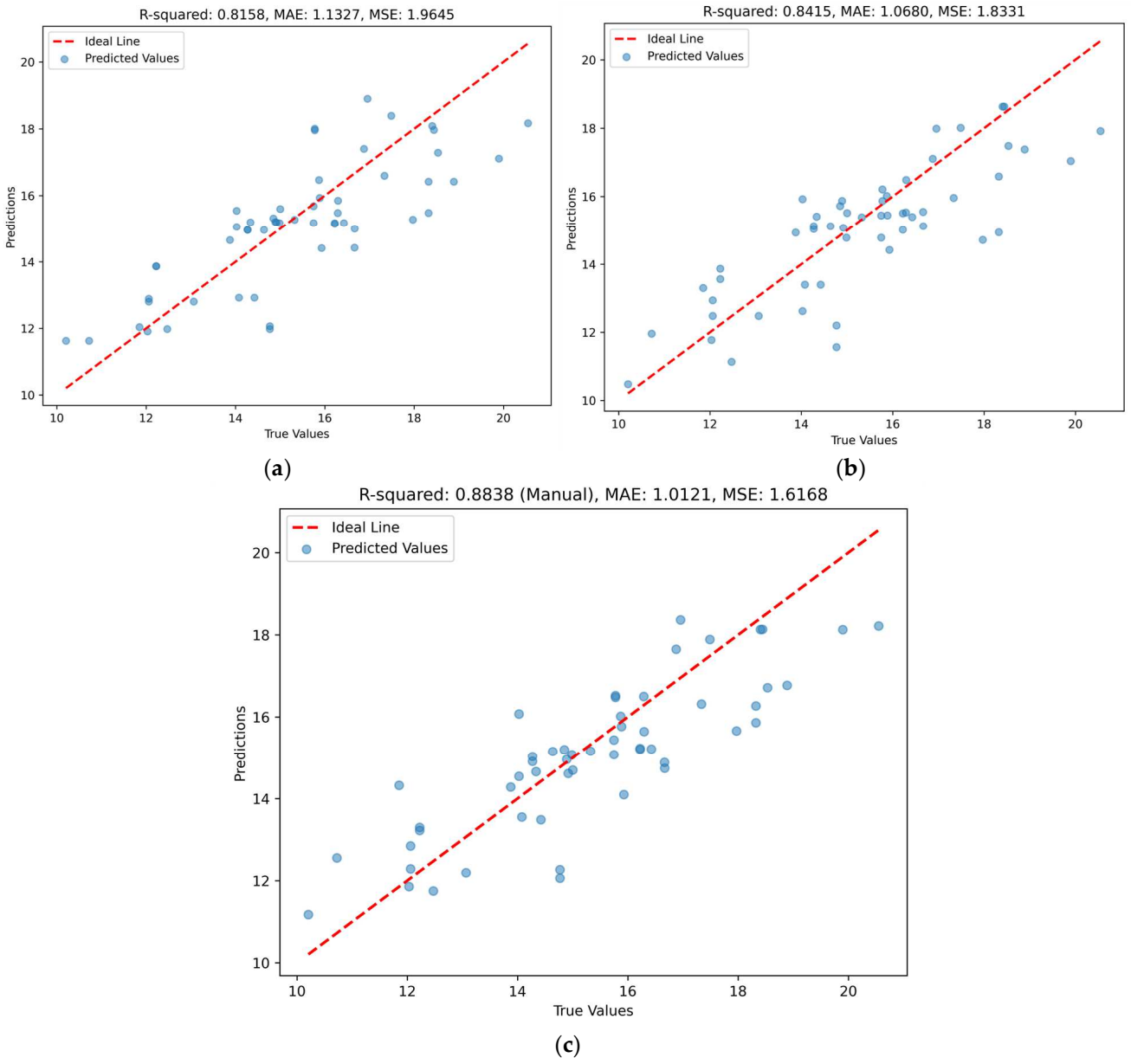


Figure 4. Full-feature band correlation algorithm result map. (**a**) GBR algorithm results. (**b**) RF algorithm results. (**c**) GBR-RF algorithm results.

The experiment involves utilizing only the sensitive single band from the previous inversion while keeping the remaining parameters unchanged. The resultant findings are illustrated in Figure 5a–i. The subsequent outcomes in this section are consolidated and presented together due to the repetitive nature of the image elements. The comparative results are presented in Figure 5j–l. An analysis of Figure 5 reveals that the results of the B6 and B8 indicators closely align with the overall band data, whereas the outcomes of the B7 band indicators exhibit a contrasting pattern due to the inversion algorithms employed. Nevertheless, the indicators of R^2, MAE, and MSE demonstrate synchronous fluctuations across all algorithms, indicating the normality of the corresponding results. Furthermore, there is an element of stability observed in the data concerning each index, underscoring the robustness and generalizability of the algorithms.

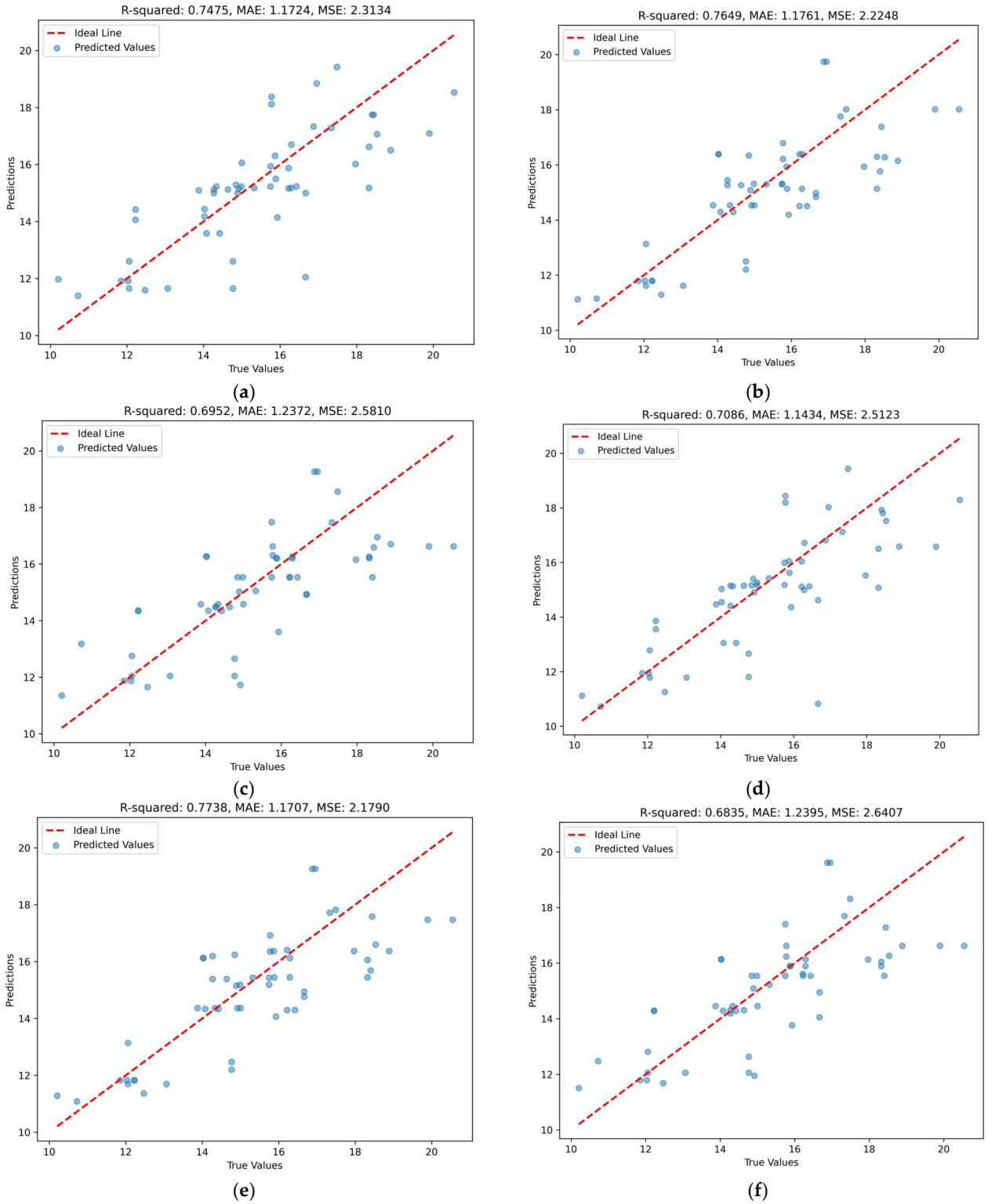


Figure 5. *Cont.*

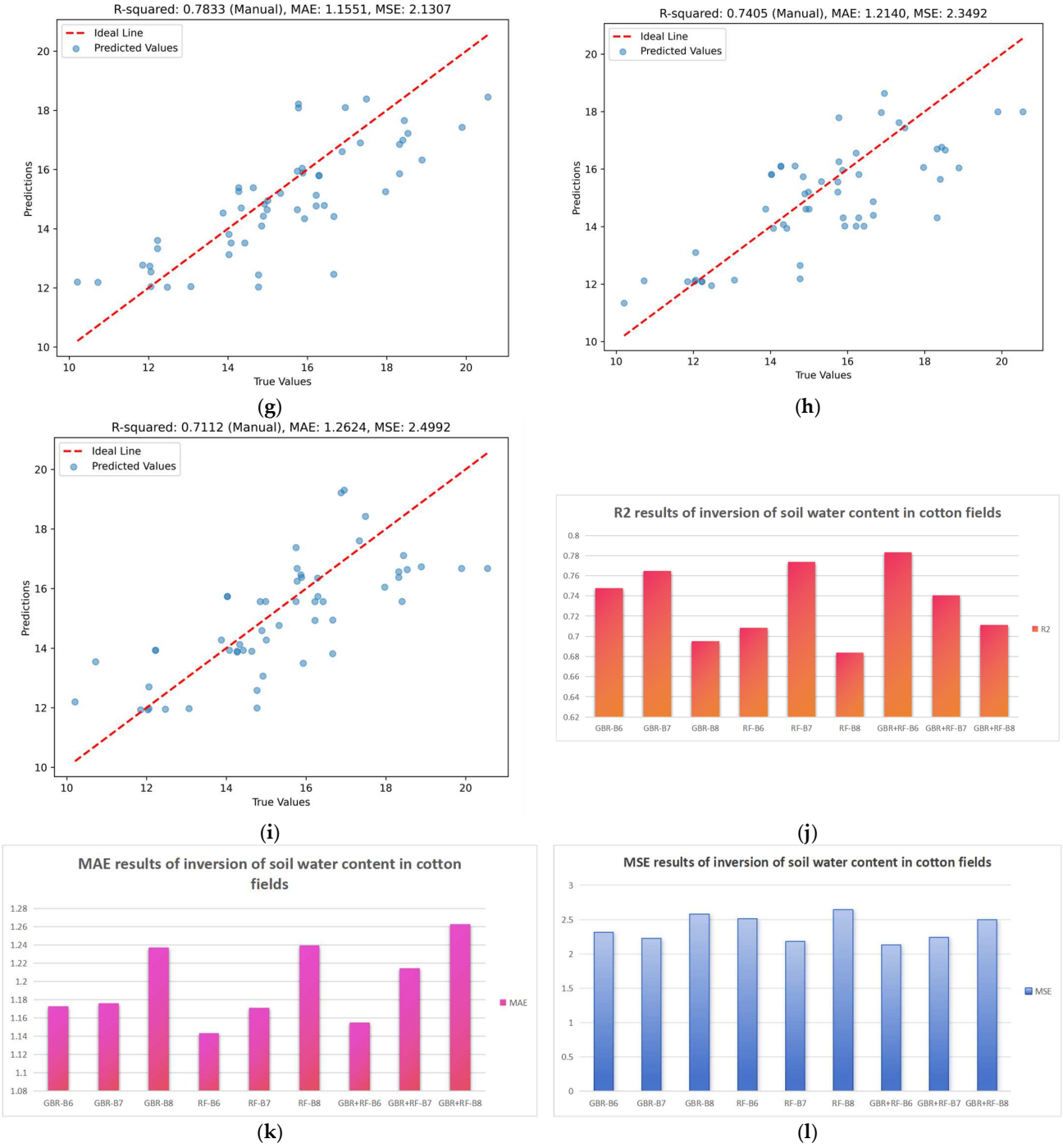


Figure 5. Results for different sensitive bands and different algorithms. (**a**) B6-band GBR results. (**b**) B7-band GBR results. (**c**) B8-band GBR results. (**d**) B6-band RF results. (**e**) B7-band RF results. (**f**) B8-band RF results. (**g**) B6-band GBR-RF results. (**h**) B7-band GBR-RF results. (**i**) B8-band GBR-RF results. (**j**) R^2 result of the inversion of soil moisture in cotton fields. (**k**) MAE result of the inversion of soil moisture in cotton fields. (**l**) MSE result of inversion of soil moisture in cotton fields.

The inversion results obtained using single-band data were found to be distinctly inferior to those achieved with full-band data. This discrepancy suggests that combining multiple bands in a certain proportion will likely yield stronger correlations and higher accuracy. However, the outcomes of inversions conducted solely with various vegetation indices (e.g., NDVI and EVI) were unsatisfactory in this study. As a result, we chose not to include vegetation indices in our experimental investigation.

4. Discussion

This study aimed to use the GBR-RF algorithm in combination with Sentinel-2 satellite spectral data to accurately estimate soil moisture content in cotton-growing areas. The goal was to provide guidance for agricultural irrigation management in order to improve water resource utilization efficiency and crop yields on farmland. In this study, the soil moisture content of a cotton field was initially measured using dedicated instruments. Subsequently, spectral images of the study area were acquired from the Google Earth Engine (GEE) platform. The collected spectral data underwent further processing in the ENVI software. Following manual data refinement, a correlation analysis was conducted to identify bands that exhibited a strong relationship with soil water content. Subsequently, a series of inversions were performed utilizing various algorithms, including gradient-boosting regression (GBR), random forest (RF), and a combination of GBR and RF, with inputs from full bands as well as the identified sensitive bands. The ensuing results were analyzed comprehensively to identify the algorithms and sensitive bands that exhibited enhanced fitting accuracy and minimized error margins. The primary aim of this process was to define the optimal algorithms and bands for improved modeling accuracy and error reduction while also strategically outlining forthcoming research trajectories. This endeavor was driven by the overarching objective of refining the association between algorithms, sensitive bands, model accuracy, and future research directions.

Recent years have seen numerous researchers combine remote sensing methods with machine learning and artificial intelligence techniques to accurately determine soil moisture content in crop fields. For instance, N. R. Prasad et al. successfully inverted cotton yields by integrating remote sensing with a crop simulation model, achieving a root mean squared error of 154 kg/ha and a correlation soil moisture content in crops coefficient R^2 of 0.46 [30], while in this study, the relevant optimal R^2 value was 0.8838, which is significantly higher than the results of the corresponding study on soil moisture in cotton. Agathos Filintas et al. conducted a study on cotton yield under various growth and irrigation conditions. The results of a two-way analysis of variance showed that the highest yield increase was 28.664%, and the highest water saving achieved was 24.941% [31]. Mireguli Ainiwaer et al. made a regional-scale estimation of soil moisture content using multi-source remote sensing parameters, achieving good results with R^2 = 0.69, RMSE = 3.48%, and RPD = 1.91 [32]. While comparing the best R^2 of 0.8838 in this study, it proved to be significantly higher than the related studies. Although these past studies are robust and detailed, the current study integrates an advanced learning algorithm for moisture inversion using Sentinel-2 satellite remote sensing data and ground soil moisture content. This approach not only enhances accuracy and robustness but also ensures greater efficiency for large-scale inversion, offering a non-destructive, swift, and highly precise method for soil moisture inversion in cotton, thereby contributing valuable insights for addressing data inversion challenges in this context.

In the experimental process, the identification of the corresponding cotton land began with field visits, followed by the collection of soil samples from the identified areas in the field. To address the significant errors in the surface data of the soil samples, inconsistent surface data were excluded. Additionally, in order to ensure consistency in depth levels, certain samples were eliminated. Subsequently, the obtained soil water content data were gathered for use in subsequent experiments. Sentinel-2 data acquisition and preprocessing were conducted on the GEE platform for the study area. During data processing, attempts were made to perform data derivative operations on the spectral band data; however, the results of these operations were unsatisfactorily compared to the original data. Consequently, the data were manually cleaned before being utilized in experiments. The data were analyzed using Pearson's coefficient, and relevant sensitive data were extracted for comparative experiments. By conducting experiments with the GBR, RF, and GBR-RF algorithms, this study achieved positive outcomes. A comparison of the results revealed that the integrated algorithm outperformed individual band inversions. Specifically, the integrated algorithm showed improvements of 12.83%, 19.35%, and 24.27% over bands B6, B7,

and B8, respectively. Furthermore, the integrated algorithm demonstrated enhancements of 8.34% and 5.03% compared to GBR and RF, respectively, highlighting the superiority of the integrated approach. Regarding the band-specific analysis, the data indicated improvements in the inversions of B6, B8, and the full band, while the results for the B7 band decreased. The combination of band data was expected to yield more accurate results due to the sensitivity of composite indices such as NDVI for vegetation being higher than that of individual bands. Although several vegetation indices were tested in the experiment, poor results were obtained, and, as a result, they were not included in the final results.

In this experiment, GBR and RF were demonstrated to be two powerful regression models capable of modeling complex nonlinear relationships. The GBR and RF algorithms are robust to the noise and outliers in input data, and they can effectively deal with the interfering factors in Sentinel-2 satellite spectral data, which improves the stability of the estimation of soil moisture content. By combining Sentinel-2 satellite spectral data, these algorithms can more accurately capture the complex relationship between soil moisture content and spectral features, thereby enhancing the prediction accuracy of soil moisture content. Additionally, Sentinel-2 satellite spectral data have a high spatial resolution and frequent observation cycles, enabling the provision of information on the spatial distribution of soil moisture content over a large area. Sentinel-2 satellite spectral data usually contain a large amount of spectral band information, leading to a high-dimensional feature space in the input data. By processing high-dimensional data and conducting feature selection and combination, effective features associated with soil moisture content were identified, thereby enhancing the generalization ability and prediction performance of the model. Moreover, the corresponding algorithms exhibit good model interpretability, allowing for the intuitive demonstration of the importance of different spectral bands in the estimation of soil moisture content. This capability is crucial for understanding the relationship between soil moisture content and spectral features, as well as for guiding subsequent agricultural management and decision-making endeavors.

In this study, there are limitations regarding the usage of the GBR and RF algorithms combined with Sentinel-2 satellite spectral data to enhance the precision and stability of soil moisture content prediction in cotton fields. One primary limitation is the relatively small size of the overall dataset, inhibiting the full exploitation of these algorithms, especially when handling the high-dimensional Sentinel-2 satellite spectral data. To ensure model generalization ability, a larger sample size is necessary. Considering non-destructive, intelligent, and efficient methodologies is essential to preventing the overcollection of data in agricultural applications. Non-destructive methods can minimize impact on crops and ecosystems, while intelligent and efficient techniques enhance agricultural production, reduce resource waste, and promote sustainable practices for long-term agricultural ecosystem health and land maintenance. By implementing various measures to minimize environmental and resource impacts, agricultural production can become more sustainable and efficient. Balancing the data collection process is crucial. Furthermore, both GBR and RF algorithms entail parameter adjustments. Such adjustments demand expertise and time, as different parameter configurations can influence model performance significantly. The mastery of the foundational concepts or the establishment of a streamlined experimental process with a specific threshold is vital to achieving optimal results. The Sentinel-2 satellite spectral data encompass details across multiple bands. The selection of input features from these bands profoundly impacts model performance, with various band combinations resulting in different predictions. While some preliminary research on this has been conducted, a deeper exploration is warranted for reasonable feature selection and optimization. Thus, while the GBR-RF algorithm alongside Sentinel-2 satellite spectral data shows advantages in predicting soil moisture content in cotton, attention should be paid to its shortcomings and limitations for improved model selection and optimization in practical scenarios. Furthermore, we should consider the generalizability and robustness of the algorithms for different crops, regions, and varieties.

The method used in this study effectively identifies the significance of various spectral bands in estimating soil moisture content, mining key features related to it, and enhancing the generalization ability and predictive performance of the model. This approach not only yields prediction results with high accuracy, stability, and broad adaptability but also offers a degree of interpretability. Consequently, the obtained results offer valuable information for enhancing agricultural production.

5. Conclusions

This study aimed to utilize relevant algorithms in combination with Sentinel-2 satellite spectral data to estimate soil moisture content in cotton cultivation areas. The ultimate goal was to provide precise soil moisture information for agricultural production, aiding in improving irrigation management and enhancing crop yields. The experiment involved processing Sentinel-2 satellite spectral data and employing the GBR-RF algorithm to successfully estimate soil moisture content in cotton cultivation areas, resulting in an R^2 value of 0.8838, an MAE of 1.0121, and an MSE of 1.6168. The research findings indicate that this method can enable high-accuracy soil moisture estimation, thus providing crucial reference information for agricultural production. Furthermore, since this study's fundamental research was conducted outside the experimental field and focused on conventionally grown cotton areas, theoretically, this method can be applied for broad-scale monitoring and estimation. Therefore, the soil moisture estimation method based on the GBR-RF algorithm and Sentinel-2 satellite spectral data is an effective technical approach, offering accurate soil moisture information for cotton cultivation areas. This research outcome provides an important scientific basis for farmland water resource management and crop production, contributing to enhancing agricultural production efficiency and sustainability.

Author Contributions: Conceptualization, X.L., J.W., J.Y. and Z.Z.; methodology, X.L. and J.W.; validation, X.L., J.W., J.Y. and Z.Z.; formal analysis, X.L., J.W., W.Z. and L.H.; investigation, J.W., Q.W., W.Z. and L.H.; resources, X.L., J.Y. and Z.Z.; data curation, J.W., Q.W., W.Z. and L.H.; writing—original draft preparation, J.W.; writing—review and editing, X.L., J.Y. and Z.Z.; visualization, J.W.; supervision, X.L., J.Y. and Z.Z.; funding acquisition, X.L., J.Y. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Oasis Ecological Agriculture Corps Key Laboratory Open Project (202002), the Corps Science and Technology Program (2021CB041, 2021BB023, and 2021DB001), Corps key areas of scientific and technological research program (2021AB022), the Tarim University Innovation Team Project (TDZKCX202306 and TDZKCX202102), and the National Natural Science Foundation of China (61563046).

Data Availability Statement: The satellite remote sensing data were obtained from GEE, and they can be obtained by visiting the corresponding official website (https://code.earthengine.google.com/ (accessed on 13 November 2023)) or by contacting the authors if they are not available. Coordinate data and soil water content data can be obtained by contacting the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bai, Y.; Mao, S.; Tian, L.; Li, L.; Dong, H. Advances and prospects of high-yielding and simplified cotton cultivation technology in Xinjiang cotton-growing area. *Sci. Agric. Sin.* **2017**, *50*, 38–50.
2. Eaton, F.M. Physiology of the cotton plant. *Annu. Rev. Plant Physiol.* **1955**, *6*, 299–328. [CrossRef]
3. Arshad, A.; Raza, M.A.; Zhang, Y.; Zhang, L.; Wang, X.; Ahmed, M.; Habib-ur-Rehman, M. Impact of climate warming on cotton growth and yields in China and Pakistan: A regional perspective. *Agriculture* **2021**, *11*, 97. [CrossRef]
4. Hsieh, Y.-L. Chemical structure and properties of cotton. In *Cotton: Science and Technology*; Woodhead Publishing: Sawston, UK, 2007; pp. 3–34.
5. Seneviratne, S.I.; Corti, T.; Davin, E.L.; Hirschi, M.; Jaeger, E.B.; Lehner, I.; Orlowsky, B.; Teuling, A.J. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Sci. Rev.* **2010**, *99*, 125–161. [CrossRef]
6. Li, N.; Lin, H.; Wang, T.; Li, Y.; Liu, Y.; Chen, X.; Hu, X. Impact of climate change on cotton growth and yields in Xinjiang, China. *Field Crops Res.* **2020**, *247*, 107590. [CrossRef]

7. Grimes, D.; Yamada, H. Relation of Cotton Growth and Yield to Minimum Leaf Water Potential 1. *Crop Sci.* **1982**, *22*, 134–139. [CrossRef]
8. Ahmad, S.; Hasanuzzaman, M. Cotton production and uses. In *Agronomy, Crop Protection, and Postharvest Technologies*; Springer Nature Singapore Pte Ltd.: Singapore, 2020.
9. Yu, L.; Gao, W.; Shamshiri, R.R.; Tao, S.; Ren, Y.; Zhang, Y.; Su, G. *Review of Research Progress on Soil Moisture Sensor Technology*; Verlag nicht ermittelbar: Beijing, China, 2021.
10. Rigden, A.; Mueller, N.; Holbrook, N.; Pillai, N.; Huybers, P. Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting US maize yields. *Nat. Food* **2020**, *1*, 127–133. [CrossRef]
11. Lang, P.; Zhang, L.; Huang, C.; Chen, J.; Kang, X.; Zhang, Z.; Tong, Q. Integrating environmental and satellite data to estimate county-level cotton yield in Xinjiang Province. *Front. Plant Sci.* **2023**, *13*, 1048479. [CrossRef]
12. Jiang, P.; Zhou, X.; Liu, T.; Guo, X.; Ma, D.; Zhang, C.; Li, Y.; Liu, S. Prediction Dynamics in Cotton Aphid Using Unmanned Aerial Vehicle Multispectral Images and Vegetation Indices. *IEEE Access* **2023**, *11*, 5908–5918. [CrossRef]
13. Leo, S.; Migliorati, M.D.A.; Nguyen, T.H.; Grace, P.R. Combining remote sensing-derived management zones and an auto-calibrated crop simulation model to determine optimal nitrogen fertilizer rates. *Agric. Syst.* **2023**, *205*, 103559. [CrossRef]
14. Wadoux, A.M.-C.; Minasny, B.; McBratney, A.B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [CrossRef]
15. Sharma, R.C. Dominant Species-Physiognomy-Ecological (DSPE) System for the Classification of Plant Ecological Communities from Remote Sensing Images. *Ecologies* **2022**, *3*, 323–335. [CrossRef]
16. Mulder, V.; De Bruin, S.; Schaepman, M.E.; Mayr, T. The use of remote sensing in soil and terrain mapping—A review. *Geoderma* **2011**, *162*, 1–19. [CrossRef]
17. Sharma, R.C. Countrywide mapping of plant ecological communities with 101 legends including land cover types for the first time at 10 m resolution through convolutional learning of satellite images. *Appl. Sci.* **2022**, *12*, 7125. [CrossRef]
18. John, K.; Abraham Isong, I.; Michael Kebonye, N.; Okon Ayito, E.; Chapman Agyeman, P.; Marcus Afu, S. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* **2020**, *9*, 487. [CrossRef]
19. de Santana, F.B.; Otani, S.K.; de Souza, A.M.; Poppi, R.J. Comparison of PLS and SVM models for soil organic matter and particle size using vis-NIR spectral libraries. *Geoderma Reg.* **2021**, *27*, e00436. [CrossRef]
20. Wang, X.; Zhang, F.; Kung, H.-T.; Johnson, V.C.; Latif, A. Extracting soil salinization information with a fractional-order filtering algorithm and grid-search support vector machine (GS-SVM) model. *Int. J. Remote Sens.* **2020**, *41*, 953–973. [CrossRef]
21. Carranza, C.; Nolet, C.; Pezij, M.; van der Ploeg, M. Root zone soil moisture estimation with Random Forest. *J. Hydrol.* **2021**, *593*, 125840. [CrossRef]
22. Zhang, Y.; Sui, B.; Shen, H.; Ouyang, L. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. *Comput. Electron. Agric.* **2019**, *160*, 23–30. [CrossRef]
23. Hosseinpour-Zarnaq, M.; Omid, M.; Sarmadian, F.; Ghasemi-Mobtaker, H. A CNN model for predicting soil properties using VIS–NIR spectral data. *Environ. Earth Sci.* **2023**, *82*, 382. [CrossRef]
24. Zhang, L.; Cai, Y.; Huang, H.; Li, A.; Yang, L.; Zhou, C. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. *Remote Sens.* **2022**, *14*, 4441. [CrossRef]
25. Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 152–170. [CrossRef]
26. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Miles, J. R-squared, adjusted R-squared. In *Encyclopedia of Statistics in Behavioral Science*; Wiley: Hoboken, NJ, USA, 2005.
29. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef]
30. Prasad, N.; Patel, N.; Danodia, A.; Manjunath, K. Comparative performance of semi-empirical based remote sensing and crop simulation model for cotton yield prediction. *Model. Earth Syst. Environ.* **2022**, *8*, 1733–1747. [CrossRef]
31. Filintas, A.; Nteskou, A.; Kourgialas, N.; Gougoulias, N.; Hatzichristou, E. A Comparison between Variable Deficit Irrigation and Farmers' Irrigation Practices under Three Fertilization Levels in Cotton Yield (*Gossypium hirsutum* L.) Using Precision Agriculture, Remote Sensing, Soil Analyses, and Crop Growth Modeling. *Water* **2022**, *14*, 2654. [CrossRef]
32. Ainiwaer, M.; Ding, J.; Kasim, N.; Wang, J.; Wang, J. Regional scale soil moisture content estimation based on multi-source remote sensing parameters. *Int. J. Remote Sens.* **2020**, *41*, 3346–3367. [CrossRef]