electronics

MDPI

Article

A Comparative Study of Machine Learning Models for Predicting Meteorological Data in Agricultural Applications

Jelena Šuljug *, Josip Spišić *, Krešimir Grgić and Drago Žagar

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, J. J. Strossmayer University of Osijek, 31000 Osijek, Croatia; kresimir.grgic@ferit.hr (K.G.)
* Correspondence: jelena.suljug@ferit.hr (J.Š.); josip.spisic@ferit.hr (J.S.)

Abstract: This study aims to address the challenges of climate change, which has led to extreme temperature events and reduced rainfall, using Internet of Things (IoT) technologies. Specifically, we monitored the effects of drought on maize crops in the Republic of Croatia. Our research involved analyzing an extensive dataset of 139,965 points of weather data collected during the summer of 2022 in different areas with 18 commercial sensor nodes using the Long-Range Wide Area Network (LoRaWAN) protocol. The measured parameters include temperature, humidity, solar irradiation, and air pressure. Newly developed maize-specific predictive models were created, taking into account the impact of urbanization on the agrometeorological parameters. We also categorized the data into urban, suburban, and rural segments to fill gaps in the existing literature. Our approach involved using 19 different regression models to analyze the data, resulting in four regional models per parameter and four general models that apply to all areas. This comprehensive analysis allowed us to select the most effective models for each area, improving the accuracy of our predictions of agrometeorological parameters and helping to optimize maize yields as weather patterns change. Our research contributes to the integration of machine learning and AI into the Internet of Things for agriculture and provides innovative solutions for predictive analytics in crop production. By focusing on solar irradiation in addition to traditional weather parameters and accounting for geographical differences, our models provide a tool to address the pressing issue of agricultural sustainability in the face of impending climate change. In addition, our results have practical implications for resource management and efficiency improvement in the agricultural sector.

Keywords: agriculture; maize; machine learning; meteorological database; artificial intelligence; support vector machine; weather forecasting; solar irradiation

check for updates

Citation: Šuljug, J.; Spišić, J.; Grgić, K.; Žagar, D. A Comparative Study of Machine Learning Models for Predicting Meteorological Data in Agricultural Applications. *Electronics* **2024**, *13*, 3284. https://doi.org/10.3390/electronics13163284

Academic Editors: Xiangjie Kong and Giovanni Pau

Received: 17 July 2024
Revised: 16 August 2024
Accepted: 17 August 2024
Published: 19 August 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/).

1. Introduction

In the field of environmental sciences, forecasting offers significant socioeconomic benefits to society at large [1]. Recent advancements in artificial intelligence (AI) have enhanced prediction capabilities, enabling the generation of numerous machine-based forecasts with improved accuracy. These advancements have been facilitated by cutting-edge technological developments [2]. Weather forecasting that includes parameters such as humidity, temperature, pressure, solar irradiation, and wind speed can be accomplished using various statistical and mathematical models. The applications of weather forecasting models developed based on data collected from ground, satellite, and radar images are extensive, covering fields such as transportation, disaster management, construction, and agriculture [3]. Additionally, these models play a critical role in optimizing agricultural practices by providing accurate and timely weather predictions [4]. To this day, the accurate prediction of hyperlocal weather parameters is still a challenging task due to the complexity of interactions among the inherent limitations in capturing fine-grained spatial and temporal variations and atmospheric parameters. By implementing IoT sensor networks, it is possible to acquire real-time data from several spatially distributed yet proximate locations

in high resolution. The increase in gathered data shall tremendously enhance the accuracy of weather parameter predictions [5].

Agriculture has long been in the spotlight in numerous countries considering the constant climate changes that affect crop yield. The costs are increasing, while the yield is decreasing, thus making farmers shift to other jobs and abandon agriculture. The Republic of Croatia is one of many countries that have different government incentives in order to increase crop production and help farmers overcome increasing production costs and the increase in extreme weather conditions. To address these issues and improve farmers' livelihoods, technology is increasingly being adopted in agriculture. Researchers and innovators are developing new techniques to enhance cultivation productivity. Weather prediction plays a crucial role in helping farmers plan crop production and estimate yields, facilitating effective crop management [6].

Embedding Internet of Things (IoT) technologies resolves various challenges across diverse applications such as smart homes, agriculture, and weather forecasting. Traditional weather reporting systems often fail to provide consistent accuracy in these increasingly adverse conditions [7]. IoT devices are capable of gathering data from proximate locations, enabling real-time monitoring of environmental parameters such as air pressure, temperature, solar irradiation, and humidity [5]. This capability is further enhanced by the deployment of spatially distributed sensors, which provide high-resolution data essential for accurate and timely weather predictions [8]. Additionally, these sensors play a crucial role in anomaly detection and improving the spatial resolution of weather models [9]. The development of the IoT has enabled the acquisition of real-time weather forecasts. This advancement allows high-precision predictive modeling to be achieved by applying machine learning (ML) algorithms to the collected data [10]. Additionally, the integration of IoT devices with ML techniques significantly enhances the accuracy and reliability of weather predictions [11]. Weather prediction encompasses a variety of methods, ranging from relatively simple environmental analyses to highly complex automated mathematical models. The temporal scope of weather forecasts can vary from one day to several months [2]. Given the non-linear relationship between crop yield and influencing factors, ML techniques are suitable for yield predictions. ML, IoT, and remote sensing technologies are transitioning traditional farming to smart farming, with applications like smart irrigation, remote monitoring, and crop growth tracking, providing innovative solutions for crop cultivation [6].

Over the past years, AI methods have gained popularity in various fields. Among these, ML stands out as a potent tool for enhancing the accuracy and reliability of models used for predicting different parameters. More precise results are often achieved when weather forecasts for smaller areas and shorter timespan are used. ML models usually have better estimation precision when they are trained using large datasets [12]. Also, it is possible to improve ML models by learning from errors [13].

Considering that there are numerous methods that can be used in the development process of weather parameter models, the idea of this manuscript was to analyze and compare various Decision Tree (DT), Support Vector Machine (SVM), Gaussian Process Regression (GPR), and linear regression methods for application in agriculture. Considering the advancements in AI and ML technologies and the overall aspiration to smart agriculture and crop cultivation, an overview of different model's efficiency is going to be presented through statistical analysis. Our research presents two major contributions. The first contribution is the comprehensive analysis of various ML technologies applied in agriculture, identifying the optimal methods for predicting specific agrometeorological parameters. Numerous studies underscore the critical role of accurate weather prediction in agriculture, significantly impacting crop yield and farm management practices. While the use of ML techniques such as linear regression, DT, SVM, and GPR for weather prediction is well documented, there is a need to identify the most effective ML techniques for specific meteorological parameters and geographical areas. Existing research often lacks comprehensive datasets essential for developing robust predictive models, a gap we

address by collecting detailed data from IoT sensors across urban, suburban, and rural areas, focusing on agricultural applications. Additionally, while previous studies have applied individual ML models to weather prediction, there is a scarcity of comparative analyses evaluating multiple models across different geographical areas and meteorological parameters. This study fills this gap by comparing 19 regression models for temperature, humidity, air pressure, and solar irradiation. Furthermore, many studies focus on broad geographical scales, potentially overlooking local weather patterns' nuances. By emphasizing predictions based on Global Positioning System (GPS) coordinates and localized data, this research aims to enhance the precision of weather forecasts for specific micro-locations, benefiting urban agriculture, which is experiencing significant growth. The proposed methods are designed to use a minimal number of input parameters, tailored to the specific parameter being estimated, with GPS coordinates playing a crucial role in determining the micro-location of the predictions. This ensures that our models are efficient and accurate for localized weather prediction. Our test results reveal that the Exponential GPR model achieved the highest R-squared (R^2) for both solar irradiation and temperature predictions. For humidity, the Exponential GPR and Bagged Trees models showed the highest accuracy. In air pressure prediction, the Rational Quadratic GPR model excelled, particularly in rural areas. These findings emphasize the robust performance of advanced regression models, especially the Exponential GPR, in accurately predicting meteorological parameters across various regions.

The second contribution is the introduction of a novel database, created using sensor data, which provides more extensive and detailed information for the selected region. This database, which consists of measured values of temperature, air pressure, solar irradiation, and humidity, supports the development of new models for similar locations in the future and offers valuable insights into urban agriculture. The detailed data captured by our sensors enable more precise agrometeorological predictions, which are essential for optimizing agricultural practices and improving crop yields in urban settings. This database stands out because it collects high-resolution weather data from a network of IoT sensors in urban, suburban, and rural areas, providing more detailed and localized information than other available databases. Unlike older databases that rely on less granular data or cover fewer locations in the region of interest, our database captures a wider range of environmental conditions with high temporal resolution, leading to more accurate and specific agrometeorological predictions. Comparative studies show that it offers superior detail and specificity compared to databases from sources like the National Centers for Environmental Information (NCEI) [14] and the European Centre for Medium-Range Weather Forecasts (ECMWF) [15]. The proposed database was used in modeling to present a comprehensive analysis of the presented data. Considering the fact that the model accuracy can be greatly affected by the size of the area taken into consideration, in part of the analysis, the data used in modeling were divided into three subcategories: rural, suburban, and urban.

Although numerous research papers describe various solutions for using IoT sensors to collect, store, and display weather data, there is a notable scarcity of open-access databases with meteorological data acquired using IoT sensors that can be used for developing ML models specifically for agricultural applications. The development of IoT-based weather reporting systems has shown the potential for creating open-access databases that provide real-time weather data [16]. Similarly, advancements in robust and affordable automatic weather stations emphasize the importance of these open-access resources for continuous and reliable data collection [17]. Furthermore, cost-effective IoT-based weather monitoring systems highlight the need for accessible databases that can enhance the precision and efficiency of weather forecasting models [18]. Existing databases such as those from the National Centers for Environmental Information [14], the European Centre for Medium-Range Weather Forecasts [15], the National Aeronautics and Space Administration (NASA) [19], the World Meteorological Organization (WMO) [20], Meteostat [21], Kaggle Datasets [22], the Global Historical Climatology Network (GHCN) [23], and OpenWeatherMap [24] are valuable resources for developing ML models. However, our objective is to create an open-

access database with a substantial amount of data for smaller, agriculturally significant areas in Croatia that are not extensively covered in many existing databases. This database will include data collected from various types of IoT sensors, enabling the analysis of model accuracy across different area sizes.

This paper is organized as follows: following the Introduction, Section 2 provides detailed information on ML technologies that can be applied for agricultural purposes, their previous application for such purposes, and an overview of technologies we used in our research. Section 3 presents the test setup, provides an overview of the hardware specification that was used for creating the proposed database and provides overall information about the database itself. Section 4 presents the developed models, along with the statistical analysis and results of model verification. Section 5 offers an overview of the test results and provides recommendations for optimal models for estimating weather parameters.

2. Modeling of Meteorological Parameters Using Machine Learning

Given that the objective of this paper is to explore the application and efficiency of ML techniques for developing models to estimate weather parameters (temperature, humidity, solar irradiation, and air pressure), the following section will provide an overview of state-of-the-art solutions documented in the literature. Emphasis will also be placed on the specific ML techniques employed in these studies and open issues.

Commonly used ML techniques include linear regression, DT, and SVM, addressing both classification and regression problems [12]. These techniques are crucial for various applications, enhancing predictive accuracy and model reliability [13]. A DT operates as a classification model, demonstrating a recursive partitioning of the instance space. Supervised ML utilizing DTs has long been applied to regression problems to enhance prognostic accuracy [25]. To achieve an optimal tradeoff between bias and variance as the models evolve from simple to complex, ensembles of trees are employed. Bagging is utilized to reduce variance [26], whereas boosting aims to mitigate errors from previous trees during data partitioning [27]. The DT approach can be employed for weather prediction by initially training the ML algorithm on historical climate data. The acquired model can then be applied to forecast various input variables such as temperature, solar irradiation, air pressure, and humidity [10]. This method has been shown to improve predictive accuracy and reliability in weather forecasting applications by leveraging past data and ML techniques [28]. Additionally, the integration of DT models with advanced ML techniques has demonstrated significant enhancements in the precision of weather predictions [29].

SVM is a powerful supervised ML algorithm with a broad range of applications, including the prediction of weather parameters. Traditionally, SVM has been employed for classification tasks. However, with the introduction of decision boundaries and hyperplanes, its use in regression tasks has increased. The objective of regression is to consider points within the decision boundary [27]. The primary goal of SVMs is to identify an optimal hyperplane that classifies data points into distinct categories. It can also accurately predict continuous target variables. While the model is being trained, the SVM algorithm can adjust hyperplane parameters to minimize mistakes in regression tasks or maximize the margin among classes [2]. SVMs are particularly effective for high-dimensional data and datasets with non-linear relationships, making them a robust ML technique [30]. These capabilities allow SVMs to provide significant improvements in prediction accuracy and model performance in various applications [31].

GPR utilizes kernel functions for non-linear regression tasks [32]. Beyond performing non-linear regression, GPR also predicts a Gaussian distribution for unfamiliar outputs [33]. By effectively employing Bayes' theorem of conditional probability, this technique interpolates observations at regular intervals [27].

Considering that this research aims to develop and compare models for estimating weather parameters for agricultural applications, specifically maize cultivation, one of the analyzed parameters was solar irradiation, which is crucial for crop growth. Solar irradiation estimation is examined in multiple research papers, given that solar energy is

extensively investigated in the context of solar power plants. Concerning this topic, notable research is detailed in [27], where solar irradiation estimation models were developed using five distinct deep learning algorithms. The study aimed to compare these methods in terms of accuracy (Root Mean Square Error (RMSE), R^2) and time complexity (prediction speed and training time) for regression tasks, with a graphical analysis of their regression training efficacy. The test results indicate that GPR achieves the highest accuracy but with increased time complexity. In contrast, ensemble methods based on DTs demonstrate faster performance but with comparatively lower accuracy. Another analysis of ML techniques, including various linear regression models, regression trees, and SVM, for modeling solar irradiation is presented in [34]. For modeling purposes, solar irradiation was estimated based on historical weather data, specifically temperature, humidity, wind speed, and air pressure. The test results indicate that the SVM with a Radial Basis Function kernel achieves the best performance compared to other methods. Additionally, the study demonstrates that solar irradiation has a strong correlation with the historical weather data utilized in the modeling process. The Pearson correlation coefficient (R) ranges between 0.75 and 1 for all four parameters, indicating a high degree of linear relationship [35].

Although there is a limited number of research papers analyzing different ML techniques specifically for agricultural applications, several studies investigate the use of ML techniques for estimating various weather parameters. A notable study is presented in [26], where the authors employ DT, SVM, random forest, and XGBoost algorithms to estimate solar irradiance and temperature. The accuracy of these methods was evaluated using absolute error (AE), mean absolute error (MAE) [36], and mean square error (MSE) [37]. The study concludes that the selection of parameters and the quality of training data significantly impact the efficiency of the proposed models, particularly when DT models are used. Although DT models demonstrate faster performance compared to other models, their efficiency is highly dependent on these factors.

One research paper employing a linear regression algorithm to estimate weather conditions is described in [38], but it does not compare the proposed solution with other ML techniques. A more comprehensive study on estimating weather conditions is presented in [39], concluding that models developed using DT achieve an accuracy of 0.82, outperforming K-Nearest Neighbor (KNN) models.

Weather forecasting using ML technologies is also investigated in [2]. Compared to previous studies, this research incorporates humidity to assess whether a day is rainy, chilly, or hot. The authors employed SVM and DT models in conjunction with artificially trained neural networks. Their findings indicate that the SVM model outperforms the DT model in terms of accuracy considering they achieve an accuracy of 50% and 80%, respectively.

In contrast to most models in the relevant literature, which rely solely on available databases, the authors in [5] integrate data from diverse sources such as traditional weather stations, user-generated reports, and IoT sensors to develop high-resolution models for estimating weather parameters. These models are designed to predict short-term, localized weather conditions. The approach combines hyperlocal weather estimation and anomaly detection using IoT sensor networks and advanced machine-learning techniques to predict wind, temperature, and precipitation. This solution is also capable of detecting weather anomalies in real-time, potentially indicating incoming extreme weather events. Our research similarly utilizes data collected from IoT sensors for hyperlocal weather estimation, aiming to predict weather changes that could adversely affect maize crops. The authors in [7] utilize IoT devices to collect real-time weather data for estimating the weather conditions in specific areas. Their approach employs Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) algorithms to improve the accuracy and simplicity of predicting humidity and temperature for localized regions. This is noteworthy as our research also involves data collection via IoT devices and focuses on comparing the accuracy of developed models across areas of varying sizes. A more cost-efficient solution is presented in [40], utilizing low-cost IoT boards and sensors to measure humidity, light, and temperature. Unlike other solutions, this approach collects data in an indoor environment

and employs a logistic regression model to estimate weather parameters in real-time. The authors note that their solution slightly outperforms other available methods. Regarding the use of ML technologies for agricultural estimations, we refer to the research presented in [6], which compared various ML techniques (Support Vector Machine, Multiple Linear Regression, and Random Forest) for estimating crop yield rates. The authors concluded that for this application, SVM and Random Forest produced similar results and outperformed the Multiple Linear Regression model in terms of accuracy.

Finally, two highly relevant papers that we want to discuss are [27,41], as they utilize open-source datasets with weather information for specific locations (Jaipur and Teknaf) in contrast to large-scale global datasets. The authors of [27] employed five deep learning techniques for weather prediction, evaluating the efficiency of the proposed algorithms using mean squared error, mean absolute error, and time complexity. Their findings indicate that DT models are the fastest but exhibit lower accuracy compared to GPR models. In contrast, the research presented in [41] uses the most comprehensive set of parameters (eight) for estimating solar irradiance, including humidity, temperature, wind speed, air pressure, precipitation, insolation clearness index, and earth skin temperature. This study was selected because it predicts solar irradiance with high accuracy using a feed-forward backpropagation neural network, despite measuring eight parameters.

The objective of our research is to achieve the best possible accuracy in predicting weather parameters while minimizing the number of input parameters measured in specific areas. By leveraging advanced ML techniques and integrating data from diverse sources, our study aims to develop robust models that require fewer input variables without compromising predictive performance. This approach not only enhances the efficiency and practicality of weather prediction models but also facilitates their application in resource-constrained environments such as small-scale agricultural operations. Through this focused methodology, we aim to provide high-precision weather forecasts that can significantly benefit agricultural productivity and decision-making processes.

For the development of models for the four meteorological parameters and analysis purposes, we used 19 different regression model types, as listed in Table 1. The models included a variety of machine-learning techniques: Linear Regression Models, regression trees, SVM, GPR models, and ensembles of trees. The model types listed in Table 1 are all available in the MATLAB regression application and generally present the most commonly used modeling technologies.

Table 1. Overview of regression model types used in the analysis.

Model Group	**Model Types**					
Linear Regression Models	Linear	Robust Linear	Interactions Linear	Stepwise Linear		
Regression Trees	Fine Tree	Medium Tree	Coarse Tree			
Support Vector Machines	Linear SVM	Quadratic SVM	Cubic SVM	Fine Gaussian	Medium Gaussian	Coarse Gaussian
Gaussian Process Regression Models	Squared Exponential	Matern 5/2 GPR	Exponential GPR	Rational Quadratic		
Ensembles of Trees	Boosted	Bagged				

When applicable, a five-fold cross-validation method was employed to mitigate overfitting by partitioning the dataset into folds and estimating accuracy for each fold. The minimum leaf size for the models was set to four. Additionally, Principal Component Analysis was disabled, and surrogate decision splits were not utilized.

3. Test Setup and Database

This section outlines an experiment conducted as part of the project titled "*An Ecosystem of Networked Devices and Services for IoT Solutions Applied in Agriculture*", which focuses on the continuous monitoring of agrometeorological and weather conditions. The experiment is critical for assessing the impact of drought on crop yields in the Republic of Croatia, where drought is the predominant cause of unprofitable yields of essential crops. Furthermore, the increasing frequency of droughts and recent climate change are expected to substantially affect the viability of strategically important crops in Croatia.

The experiment was carried out from 13 July 2022 to 29 September 2022, in the regions of Osijek and Tovarnik. It employed 18 commercial sensor nodes utilizing the Long-Range Wide Area Network (LoRaWAN) protocol to transmit data to the LORIOT server. Subsequently, these data were relayed to the university server's database via a web socket. The devices were mounted on tripods to facilitate the real-time collection of meteorological data. The sensors used in this study included the METEOHELIX® IoT PRO [42], LoRaWAN Weather Station (WS100LRW/LW) [43], and ELEVEN PARAMETER WEATHER STATION FOR LoRaWAN® [44]. Detailed technical specifications are provided in Table 2. The locations of 15 sensor nodes employed in and around Osijek are shown in Figure 1. The remaining three nodes are placed in Tovarnik, Croatia. The sensor locations were strategically chosen to cover rural, suburban, and urban areas, with a particular emphasis on urban areas due to the recent surge in urban agriculture. This focus on urban settings reflects the growing interest and expansion in urban farming practices. However, there were also technological constraints related to network connectivity (the distance from the communication tower) that influenced the placement of sensors. Ensuring reliable connectivity was essential for real-time data collection and monitoring, and this requirement sometimes limited the ability to deploy sensors in more remote rural locations. Despite these challenges, the deployment aimed to provide a comprehensive dataset that captures the diverse environmental conditions across different geographical contexts.

Table 2. Overview of specifications of sensors used for gathering meteorological data.

Station	Temperature Accuracy	Humidity Accuracy	Pressure Accuracy	Solar Radiation Accuracy	Wireless Technology
MeteoHelix IoT Pro [42]	±0.1 °C 0–65 °C	±1.5%	±0.5 hPa	5% of the daily total	LoRaWAN
UBIQ Weather Station [43]	40.15–64.85 °C	0–100 RH%	N/A	0–1800 W/m^2	LoRaWAN
Decentlab DL-ATM41 [44]	±0.6 °C	±3% RH	±0.1 kPa	±5%	LoRaWAN

During the experiment, the frequency of data collection was once every 10 min, thus a total of 139,965 records of weather conditions were collected from the field and surrounding areas, specifically the cities of Osijek and Tovarnik. These records were categorized into three groups based on their geographical locations: urban, suburban, and rural. Six sensors were deployed in urban areas (Osijek), nine in suburban areas (near Osijek), and three in rural areas (Tovarnik). The collected data included meteorological parameters such as temperature, air humidity, solar irradiation, and air pressure. These data were processed and analyzed to accurately reflect the agrometeorological and physiological conditions. The comprehensive database generated from this experiment is made available through a GitHub link provided at the conclusion of this paper, facilitating access and utilization by researchers in the field. The data are crucial for agricultural production in the Republic of Croatia as they enable precise analysis of the impact of weather conditions, including drought and other extreme events, on crop yields and agricultural viability. Additionally, the findings from this study highlight the potential of IoT solutions to drive the development of new technologies and methodologies that contribute to the sustainability and productivity of agricultural practices. This experiment underscores the significant

benefits of IoT applications in agriculture, demonstrating the feasibility of continuous monitoring and real-time analysis of agrometeorological and physiological conditions of crops. The flowchart provided in Figure 2 outlines the process of developing machine-learning models for meteorological parameter estimation. Models were developed for four meteorological parameters: temperature, air pressure, humidity, and solar irradiation. These models were constructed separately for urban, suburban, and rural areas, as well as for the aggregated data representing the Slavonia region in Croatia. Following data collection, model training is conducted for each parameter using 19 different regression models (Table 1), each incorporating various numbers of input parameters to explore their effects on model performance. The selection of the 19 different regression models is based on their performance in previous studies (Section 2) and their suitability for the data characteristics. The modeling process involved training various ML techniques, including Linear Regression Models, DT, SVM, and GPR models, with carefully selected parameters. For Linear Regression Models, the parameters focused on ordinary least squares fitting. DTs were configured with a minimum leaf size of four and no surrogate decision splits, ensuring a balance between model complexity and interpretability. SVM models utilized different kernel functions, such as linear, quadratic, and cubic, to capture non-linear relationships in the data. GPR models employed kernel functions like squared exponential and rational quadratic to accommodate the non-linearity and variability in meteorological parameters. Hyperparameter selection was refined through a five-fold cross-validation process, which mitigated overfitting and ensured model robustness. By partitioning the dataset into five subsets, each model's performance was comprehensively evaluated across different data segments. The data from 14 sensor nodes were utilized in the modeling process. After initial model training, the process involves the selection of optimal input parameters based on key performance metrics: RMSE, R^2, prediction speed, and training time. This step ensures that the models use the most effective and efficient set of input variables. Subsequently, an analysis and selection of four models per parameter for each area is performed, again using RMSE, R^2, prediction speed, and training time to determine the best candidates. The best model is chosen based on the highest R^2 value, the lowest RMSE, and the shortest training and prediction times. The next step involves the selection of one optimal model per meteorological parameter for each geographical area, refining the choices to the single best-performing model for each parameter and area type. Validation using a subset of the data is then conducted to confirm the generalizability and accuracy of these selected models. For validation using a subset of the data, we gathered data from an additional four sensor nodes. Specifically, two sensor nodes from urban areas, one from a suburban area, and one from a rural area were designated exclusively for testing purposes. The sensors selected for verification were strategically chosen based on their GPS locations to present a challenging task for the proposed models. This careful selection ensures that the models are rigorously tested under diverse and demanding conditions, thereby validating their robustness and accuracy. By encompassing a wide range of environmental variables and micro-locations, the verification process provides a comprehensive assessment of the models' performance, ensuring their reliability and effectiveness in real-world applications. To facilitate a comprehensive understanding of the proposed database, Figure 3 illustrates data collected for urban areas that were used in modeling, and Figure 4 illustrates data collected for urban areas that were used for verification. Figures 3 and 4 present scatter plots illustrating the measured values of temperature, humidity, solar irradiation, and pressure across a span of 2.5 months. The data are depicted for each daytime hour, with multiple measurements captured, reflecting the variability of the environmental conditions over the observation period.



Figure 1. Locations of 15 sensor nodes employed in and around Osijek, Croatia.

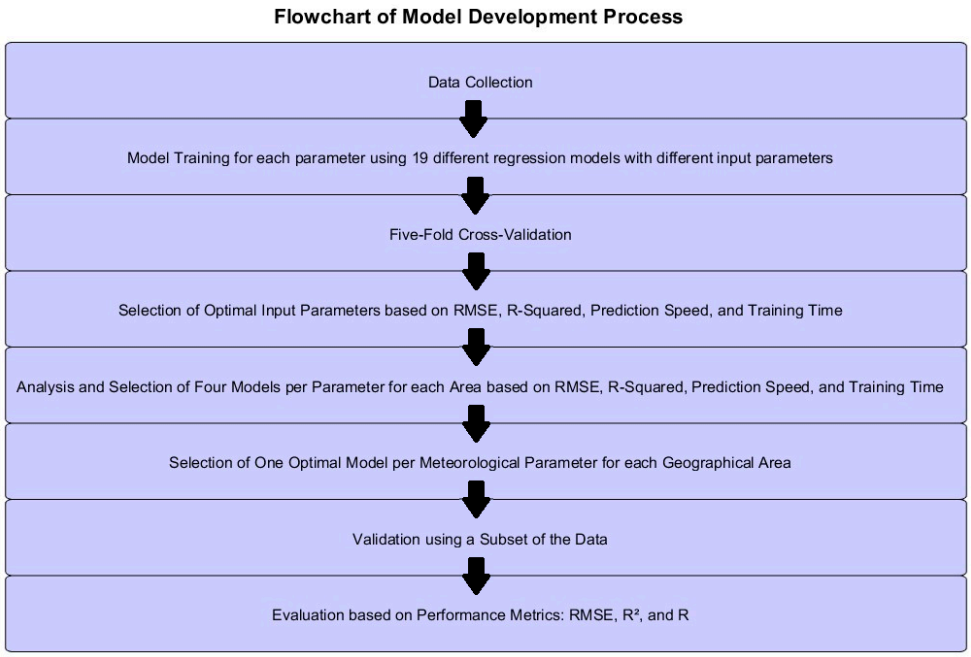


Figure 2. Flowchart of model development process.

Finally, the modeling process ends with Evaluation Based on Performance Metrics, including RMSE, R^2, and R, to ensure they meet the required standards of accuracy and reliability for practical application. The selected models for each parameter and geographical area were chosen based on their performance in terms of RMSE, R^2, prediction speed, and training time. The Rational Quadratic GPR and Exponential GPR models consistently showed high accuracy, low RMSE, and high R^2 and were therefore chosen as the best models for predicting meteorological parameters across various regions. The detailed evaluation ensures that the selected models provide reliable and efficient predictions, making them highly suitable for practical applications in weather forecasting and agricultural planning. This structured and iterative approach ensures that the developed models are both robust and efficient, capable of providing accurate meteorological predictions for different geographical areas.

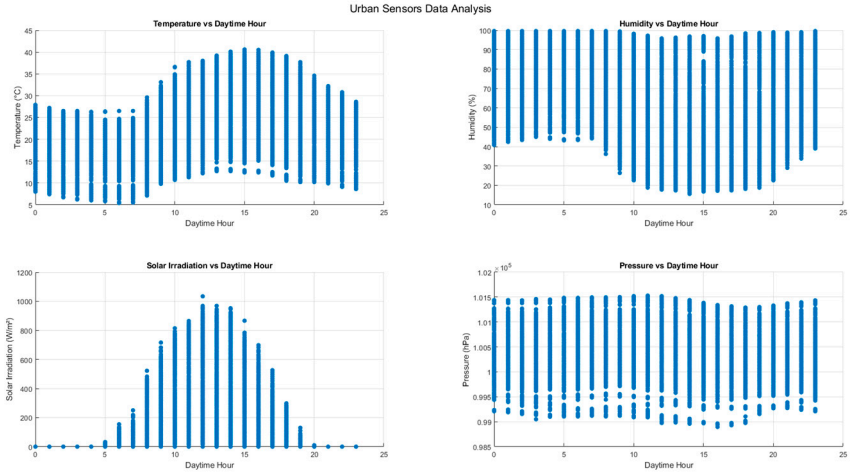


Figure 3. Weather data collected for urban areas that were used in modeling.

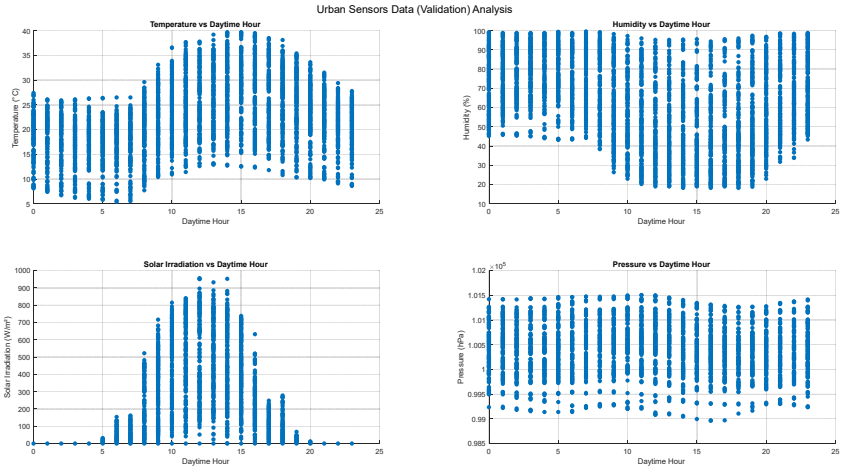


Figure 4. Weather data collected for urban areas that were used for verification.

4. Results

After generating 304 distinct models—19 for each meteorological parameter across various regional scales—an additional 380 models were created to determine the minimal-yet-sufficient number of input parameters required for accurately estimating each meteorological parameter. Pressure was modeled in three test cases that differed in the number of input parameters: the first test case included latitude, longitude, month, hour, temperature, and humidity; the second was without humidity, and the third was without temperature. Solar irradiation and humidity were modeled in two test cases that differed in the number of input parameters: the first test case included latitude, longitude, month, hour, and temperature, and the second was without temperature. Temperature was modeled in two test cases that differed in the number of input parameters: the first test case included latitude, longitude, month, hour, and humidity, and the second was without humidity. The developed models were evaluated using several performance metrics: Root Mean Square Error, R^2, mean squared error, mean absolute error, prediction speed, and training time. Given the extensive data collected for all 684 models, all test results are available alongside the proposed database [45].

Based on the analysis of the results, the optimal input parameters for predicting air pressure were identified as latitude, longitude, month, hour, temperature, and humidity. For predicting temperature, the input parameters were determined to be latitude, longitude, month, hour, and humidity. For solar irradiation and humidity predictions, the input parameters were identified as latitude, longitude, month, hour, and temperature.

In the second stage of the analysis, four models per meteorological parameter for each area were identified as having the most promising performance, resulting in a total of 64 models. The codes of these models are available for testing alongside the proposed database. The R^2 values that were selected to be used as a measure of model accuracy for all models are presented in Table 3. The analysis indicates that regression trees, GPR models, and ensembles of trees achieve the highest R^2 values, thus providing the highest accuracy for modeling meteorological data collected via IoT sensor nodes. These models are particularly effective for agricultural applications.

Table 3. R^2 values for 64 models generated in the second stage of analysis.

Parameter	Model Type	All Nodes	Urban Area	Suburban Area	Rural Area
Pressure	Fine Tree	0.61	0.63	0.58	0.55
	Fine Gaussian SVM	0.53	0.56	0.51	0.5
	Bagged Trees	0.61	0.6	0.57	0.41
	Rational Quadratic GPR	0.72	0.77	0.72	0.79
Solar irradiation	Medium Tree	0.87	0.85	0.87	0.88
	Fine Gaussian SVM	0.84	0.83	0.86	0.89
	Bagged Trees	0.87	0.84	0.87	0.86
	Exponential GPR	0.88	0.86	0.89	0.90
Humidity	Medium Tree	0.87	0.85	0.87	0.86
	Fine Gaussian SVM	0.86	0.84	0.87	0.86
	Bagged Trees	0.88	0.85	0.88	0.82
	Exponential GPR	0.88	0.86	0.88	0.88
Temperature	Coarse Tree	0.87	0.86	0.87	0.83
	Fine Gaussian SVM	0.86	0.85	0.86	0.83
	Bagged Trees	0.87	0.86	0.87	0.8
	Exponential GPR	0.88	0.87	0.88	0.84

The analysis reveals that regression trees, GPR models, and ensembles of trees exhibit the highest R^2 values, indicating superior accuracy in modeling meteorological data collected via IoT sensor nodes for agricultural purposes. For air pressure, the Rational Quadratic GPR model demonstrated the highest R^2 values, particularly in the case of rural areas, with values reaching 0.79. Bagged Trees and Fine Tree models also showed strong performance across all areas, though not as high as the GPR models. For the solar irradiation parameter, the Exponential GPR model consistently achieved the highest R^2 values across all areas, with values reaching up to 0.90 in rural areas and 0.88 overall. The Medium Tree and Bagged Trees models also performed well, with R^2 values ranging from 0.85 to 0.88 across different regions. Specifically, the Medium Tree model achieved an R^2 value of 0.88 in rural areas, while the Bagged Trees model had values close to 0.87–0.88 in suburban and rural areas. For humidity, both the Exponential GPR and Bagged Trees models exhibited high R^2 values around 0.88, indicating high accuracy. Medium Tree and Fine Gaussian SVM models also performed effectively, though slightly lower than the GPR and Bagged Trees models. Regarding temperature, the Exponential GPR model again led with the highest R^2 values, up to 0.88. Coarse Tree and Bagged Trees models showed solid performance with R^2 values around 0.87.

The RMSE values for solar irradiation models in our study show considerable variation across different geographical areas and model types, reflecting their performance and accuracy. For the Medium Tree model, the RMSE values are 82.129 W/m^2 (0.0821 kW/m^2) for all sensors, 75.806 W/m^2 (0.0758 kW/m^2) for urban areas, 86.111 W/m^2 (0.0861 kW/m^2) for

suburban areas, and 88.902 W/m^2 (0.0889 kW/m^2) for rural areas. The Fine Gaussian SVM recorded RMSE values of 91.895 W/m^2 (0.0919 kW/m^2) for all sensors, 80.888 W/m^2 (0.0809 kW/m^2) for urban areas, 87.623 W/m^2 (0.0876 kW/m^2) for suburban areas, and 87.029 W/m^2 (0.0870 kW/m^2) for rural areas. The Bagged Trees model demonstrated RMSE values of 81.206 W/m^2 (0.0812 kW/m^2) for all sensors, 78.845 W/m^2 (0.0788 kW/m^2) for urban areas, 85.572 W/m^2 (0.0856 kW/m^2) for suburban areas, and 96.037 W/m^2 (0.0960 kW/m^2) for rural areas. Notably, the Exponential GPR model achieved the lowest RMSE values, with 77.990 W/m^2 (0.0780 kW/m^2) for all sensors, 72.365 W/m^2 (0.0724 kW/m^2) for urban areas, 79.079 W/m^2 (0.0791 kW/m^2) for suburban areas, and 82.773 W/m^2 (0.0828 kW/m^2) for rural areas.

In comparison to other studies, our results align well with reported RMSE values for solar irradiation predictions using ML models, which range from 40.87 W/m^2 to 94.89 W/m^2 (0.0409 to 0.0949 kW/m^2) [46]. A relevant study presented in Table 1 of [47] reported RMSE values ranging from 75.23 W/m^2 to 146.22 W/m^2 (0.0752 to 0.1462 kW/m^2), which are comparable to our results, indicating the robustness and effectiveness of our selected models. The urban areas in our study exhibited the lowest RMSE values, indicating the highest accuracy for solar irradiation predictions. This is consistent with other research findings that suggest urban areas benefit from more stable and predictable environmental conditions compared to rural and suburban areas, leading to more accurate model predictions.

For all sensors combined, the Fine Tree model for pressure achieved an RMSE of 2.8461 hPa, indicating moderate accuracy with high processing efficiency. The Fine Gaussian SVM model recorded an RMSE of 3.1398 hPa, balancing accuracy and resource use. The Bagged Trees model had an RMSE of 2.8417 hPa, showing robust performance. The Rational Quadratic GPR model achieved the lowest RMSE value of 2.4369 hPa, indicating superior accuracy despite a more resource-intensive process. In urban areas, the Fine Tree model's RMSE was 2.7039 hPa, demonstrating good accuracy. The Fine Gaussian SVM recorded an RMSE of 2.9425 hPa with moderate prediction speed. The Bagged Trees model showed an RMSE of 2.8352 hPa, making it efficient for urban data processing. The Rational Quadratic GPR model achieved the lowest RMSE of 2.1217 hPa, excelling in accuracy in urban environments. For suburban areas, the Fine Tree model recorded an RMSE of 2.9488 hPa, while the Fine Gaussian SVM had an RMSE of 3.1751 hPa. The Bagged Trees model had an RMSE of 2.9936 hPa. The Rational Quadratic GPR model again demonstrated the lowest RMSE of 2.4009 hPa, indicating its superior performance. In rural areas, the Fine Tree model's RMSE was 3.3939 hPa. The Fine Gaussian SVM recorded an RMSE of 3.5445 hPa. The Bagged Trees model showed an RMSE of 3.8824 hPa. The Exponential GPR model achieved the lowest RMSE of 2.3321 hPa, highlighting its effectiveness in rural settings.

For all sensors combined, the Medium Tree model for humidity achieved an RMSE of 8.1906%, reflecting moderate accuracy with high computational efficiency. The Fine Gaussian SVM model reported an RMSE of 8.3525%, balancing precision and resource utilization. The Bagged Trees model had an RMSE of 7.8979%, indicating robust performance. The Exponential GPR model attained the lowest RMSE value of 7.8392%, showcasing superior accuracy despite being more resource intensive. In urban areas, the Medium Tree model's RMSE was 8.5452%, demonstrating good accuracy. The Fine Gaussian SVM recorded an RMSE of 8.8605%, with moderate prediction speed. The Bagged Trees model exhibited an RMSE of 8.4532%, making it efficient for urban data processing. The Exponential GPR model achieved the lowest RMSE of 8.2139%, excelling in accuracy in urban settings. For suburban areas, the Medium Tree model recorded an RMSE of 8.1297%, while the Fine Gaussian SVM had an RMSE of 8.2496%. The Bagged Trees model had an RMSE of 7.9374%. The Exponential GPR model again demonstrated the lowest RMSE of 7.785%, indicating its superior performance. In rural areas, the Medium Tree model's RMSE was 8.2669%. The Fine Gaussian SVM recorded an RMSE of 8.2611%. The Boosted Trees model showed an RMSE of 9.4674%. The Exponential GPR model achieved the lowest RMSE of 7.814%,

highlighting its effectiveness in rural environments. Overall, suburban areas in our study exhibited the lowest RMSE values, indicating the highest accuracy for humidity predictions. These results underscore the robustness and precision of the Exponential GPR model, making it a valuable tool for environmental monitoring and forecasting applications.

When comparing our results with other relevant research, studies on humidity prediction using ML report RMSE values ranging from 5% to 10%, depending on model complexity and the dataset used. For instance, a study utilizing LSTM and ANFIS models for daily relative humidity forecasting in Turkey reported RMSE values of 5.95% to 7.67% across various provinces [48]. This comparison indicates that our models, particularly the Exponential GPR model, perform competitively with state-of-the-art models, confirming their robustness and reliability in predicting humidity across various environmental conditions.

For all sensors combined, the Coarse Tree model for temperature had an RMSE of 2.3817 °C, showing decent accuracy with good efficiency. The Fine Gaussian SVM model had an RMSE of 2.4075 °C, balancing accuracy and resource use well. The Bagged Trees model recorded an RMSE of 2.2997 °C, indicating solid performance. The Exponential GPR model stood out with the lowest RMSE of 2.2805 °C, despite needing more resources. In urban areas, the Medium Tree model's RMSE was 2.4158 °C, showing good accuracy. The Fine Gaussian SVM had an RMSE of 2.4548 °C, with moderate prediction speed. The Bagged Trees model had an RMSE of 2.3817 °C, making it efficient for urban data. The Exponential GPR model achieved the lowest RMSE of 2.3230 °C, excelling in accuracy in urban settings. For suburban areas, the Coarse Tree model had an RMSE of 2.3590 °C, while the Fine Gaussian SVM recorded an RMSE of 2.3902 °C. The Bagged Trees model had an RMSE of 2.3186 °C. The Exponential GPR model again had the lowest RMSE of 2.2801 °C, indicating its strong performance. In rural areas, the Medium Tree model's RMSE was 2.8438 °C. The Fine Gaussian SVM recorded an RMSE of 2.8620 °C. The Boosted Trees model had an RMSE of 3.0385 °C. The Exponential GPR model achieved the lowest RMSE of 2.7016 °C, showing its effectiveness in rural settings.

When comparing our results with other relevant research, studies on temperature prediction using ML report RMSE values ranging from 0.5 °C to 3 °C, depending on model complexity and the dataset used. For instance, a study focusing on temperature forecasting using ML models like LSTM reported RMSE values around 2.3 °C, indicating high accuracy in daily temperature predictions [49]. Furthermore, a paper on ultra-low-temperature measurement using an SSA-PSO-ELM network model reported an RMSE of 3.3081 °C for SVR and 4.4835 °C for the least squares method, showcasing the potential for higher RMSE values in specific conditions [50]. This comparison underscores the effectiveness of our approach, with our RMSE values aligning well with those reported in the literature.

Overall, these findings underscore the effectiveness of advanced regression models, such as GPR and ensembles of trees, in capturing the complexities of meteorological data for agricultural applications. These models provide important tools for real-time monitoring and analysis, enhancing the precision of agricultural forecasts and decision-making processes. The comprehensive database, along with these high-performing models, offers valuable resources for ongoing research and practical implementations in the field of smart agriculture.

For predicting solar irradiation, temperature, air pressure, and humidity, tree-based models like Fine Trees and Bagged Trees are often recommended due to their reliable performance and robustness. As can be seen from the results, SVMs are also worth considering, especially for temperature and pressure predictions, if computational resources are carefully managed. GPR models, including variants like Squared Exponential GPR, Matern 5/2 GPR, Exponential GPR, and Rational Quadratic GPR, are highly effective for regression tasks, including predicting weather parameters. While using tree models for their lower computational cost and good scalability in large-scale applications can be beneficial, our results show that GPR models, particularly Exponential GPR, perform strongly when

computational resources are available and high accuracy is needed, especially in cases of smaller datasets.

This study also provides comprehensive data on the prediction speed and training time for various regression models used to predict meteorological parameters, highlighting significant differences in computational efficiency. For instance, in predicting air pressure, the Fine Tree model achieved a high prediction speed of 1,500,000 observations per second with a relatively short training time of 40.935 s, whereas the Rational Quadratic GPR model, despite its higher accuracy, had a much lower prediction speed of 700 observations per second and a substantially longer training time of 12,094 s. Similarly, for solar irradiation, the Medium Tree model exhibited an impressive prediction speed of 2,000,000 observations per second and a training time of 28.872 s, while the Exponential GPR model, known for its accuracy, lagged with a prediction speed of 770 observations per second and a training time of 10,791 s. In humidity prediction, the Medium Tree model again outperformed in speed with 2,000,000 observations per second and a training time of 7.7101 s, compared to the Exponential GPR model's 1600 observations per second and 3871.3 s of training time. For temperature, the Coarse Tree model led with the fastest prediction speed of 2,100,000 observations per second and the shortest training time of 1.8304 s, while the Exponential GPR model had the lowest prediction speed of 1000 observations per second and a longer training time of 5913.1 s. These results indicate that more complex models like the Exponential GPR and Rational Quadratic GPR models offer the highest accuracy, which is what we aimed for in this study, but at the cost of using more computational resources and time. The aforementioned results suggest that while more sophisticated models may provide better accuracy, their slower prediction speeds and longer training times may limit their practicality in some real-time or resource-constrained environments, making it crucial to balance model complexity with computational efficiency depending on the application's needs.

In the subsequent phase of the research, one model per meteorological parameter for each geographical area was selected, resulting in a total of 16 best-performing models. The Rational Quadratic GPR model was identified as the best-performing model for air pressure in suburban and urban areas, as well as when all data were used in the modeling process. In the case of rural areas, the Exponential GPR model was selected as the best-performing model for air pressure. Furthermore, Exponential GPR models were also chosen as the best-performing models for solar irradiation, humidity, and temperature across all area sizes. These selections underscore the effectiveness of GPR models in accurately capturing the complexities of meteorological data across various environmental conditions.

Data including RMSE, R^2, prediction speed, and training time are presented in Table 4. Table 4 presents an evaluation of regression models for predicting temperature, humidity, solar irradiation, and pressure across different geographical areas—urban, suburban, rural, and all data combined. The models generally demonstrate low RMSE values, indicating accurate predictions, and high R^2 values, showing a strong fit between predicted and actual values. Models for rural areas exhibit the highest R^2 values for solar irradiation and pressure, indicating robust performance in these contexts. Prediction speed varies significantly, with rural models displaying much higher speeds, particularly for humidity and temperature, compared to models using all data, which have notably lower speeds for solar irradiation and pressure. Training time also varies, with models trained on all data requiring significantly more time, especially for solar irradiation and pressure, reflecting their resource-intensive nature. In contrast, rural models have much shorter training times, indicating greater efficiency. Urban and suburban models ensure a balance between accuracy and efficiency, offering reasonable prediction speeds and moderate training times, making them practical for real-time applications.

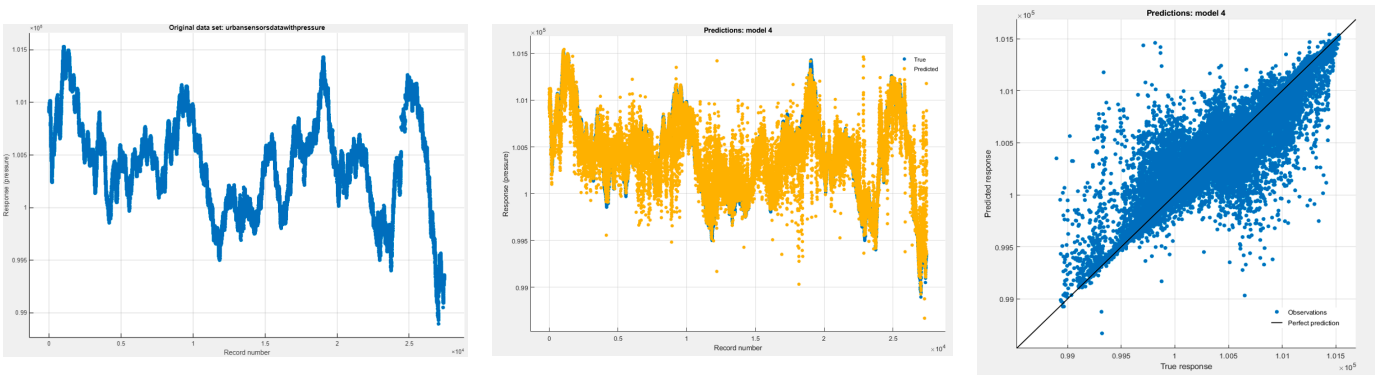
Overall, the table highlights the trade-offs between accuracy, as indicated via RMSE and R^2, and computational, i.e., time efficiency, as shown with prediction speed and training time. Rural models are more efficient but slightly less accurate, while models using all data are more accurate but demand more computational resources. Urban and suburban models

provide a middle ground, balancing accuracy and efficiency, which is crucial for selecting the appropriate model based on specific application needs, whether prioritizing accuracy, speed, or computational efficiency.

Table 4. RMSE, R^2, prediction speed, and training time values for all analyzed areas and parameters.

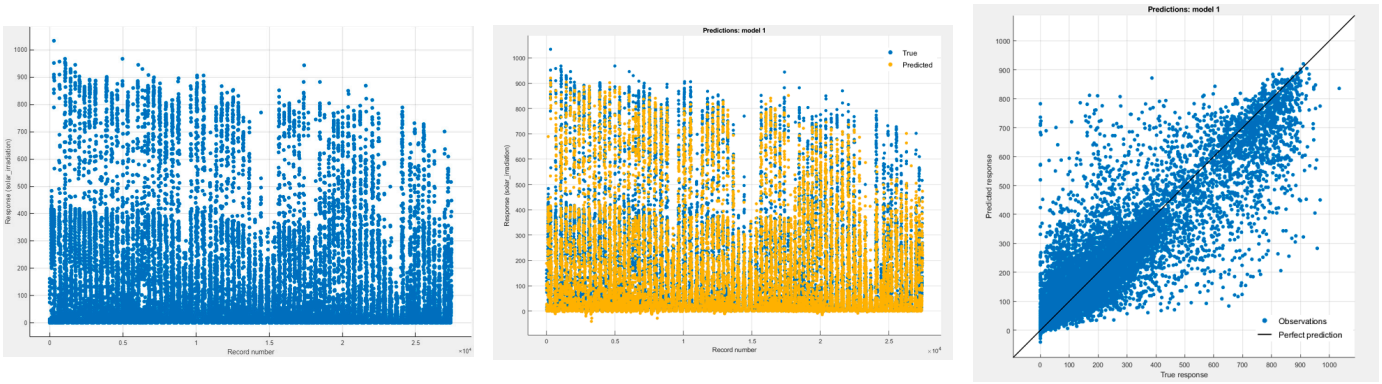
Area size	Measure	Temperature	Humidity	Solar Irradiation	Pressure
All data	RMSE	2.280	7.8392	77.99	2.437
	R^2	0.88	0.88	0.88	0.72
	Prediction speed [obs/s]	1000	1600	770	700
	Training time [s]	5913.1	3871.3	10,791	12,094
Urban area	RMSE	2.32	8.2139	72.365	2.122
	R^2	0.87	0.86	0.86	0.77
	Prediction speed [obs/s]	1600	2900	4200	2300
	Training time [s]	778.37	695.62	528.83	1729.1
Suburban area	RMSE	2.280	7.785	79.079	2.401
	R^2	0.88	0.88	0.89	0.72
	Prediction speed [obs/s]	1600	2200	1700	1200
	Training time [s]	1878	1505	1887.3	3970.6
Rural area	RMSE	2.702	7.814	82.773	2.332
	R^2	0.84	0.88	0.9	0.79
	Prediction speed [obs/s]	6500	11,000	7600	7900
	Training time [s]	122.94	82.044	113.65	143.07

Figures 5–8 depict response plots and predicted vs. actual data plots for urban area models. In the verification phase of the modeling process, the performance metrics for temperature, humidity, solar irradiation, and pressure were further analyzed across different geographical areas: urban, suburban, and rural. The models were evaluated using the Pearson correlation coefficient and R^2 values to assess their accuracy and reliability (Table 5). For temperature, when considering all data, the correlation coefficient was 0.92, with an R^2 value of 0.85, indicating a strong relationship between the predicted and actual temperature values.



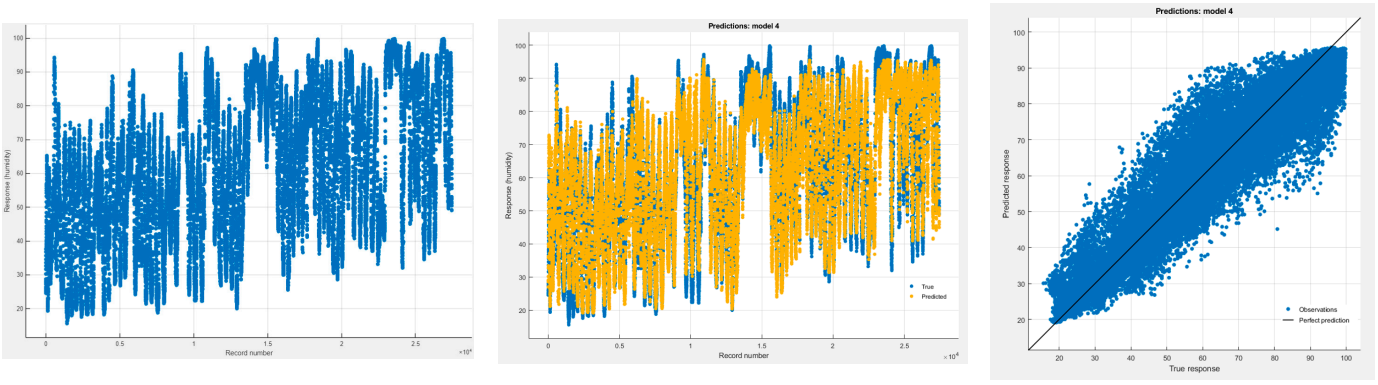
(**a**) (**b**) (**c**)

Figure 5. (**a**) Air pressure response plot; (**b**) air pressure true values; and (**c**) air pressure predicted vs. actual data plot.



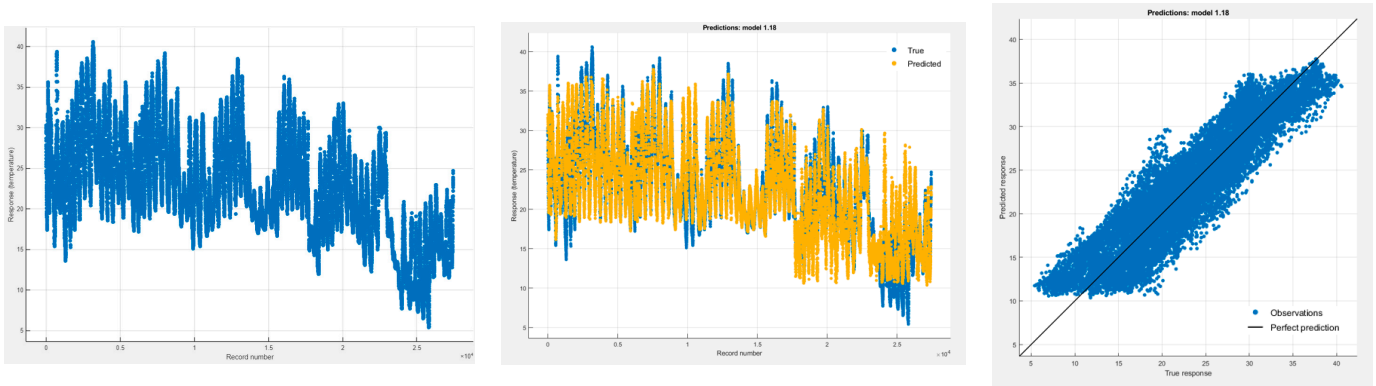
(**a**) (**b**) (**c**)

Figure 6. (**a**) Solar irradiation response plot; (**b**) solar irradiation true values; and (**c**) solar irradiation predicted vs. actual data plot.



(**a**) (**b**) (**c**)

Figure 7. (**a**) Humidity true values; (**b**) humidity response plot; and (**c**) humidity predicted vs. actual data plot.



(**a**) (**b**) (**c**)

Figure 8. (**a**) Temperature true values; (**b**) temperature response plot; and (**c**) temperature predicted vs. actual data plot.

In urban areas, the models achieved the highest R-value of 0.93 and an R^2 value of 0.87, reflecting excellent model performance. Suburban areas also demonstrated high accuracy with R = 0.94 and R^2 = 0.88. However, in rural areas, the R-value was lower at 0.77, and the R^2 value was 0.59, suggesting that the model was less accurate in these areas compared to urban and suburban areas. This can be explained by a lower amount of data used in the modeling process that covers a larger area and is something to be improved in the future, although the predictions are still considered to be adequate.

Table 5. Correlation coefficient (R) and R^2 values for all analyzed areas and parameters.

Area size	Measure	Temperature	Humidity	Solar Irradiation	Pressure
All data	R	0.9234	0.9291	0.9115	0.8099
	R^2	0.8512	0.8602	0.8285	0.6442
Urban area	R	0.9338	0.9479	0.9654	0.9770
	R^2	0.8719	0.8982	0.9316	0.9501
Suburban area	R	0.9373	0.9471	0.9547	0.9177
	R^2	0.8781	0.8968	0.9064	0.8361
Rural area	R	0.7702	0.8692	0.8746	0.8751
	R^2	0.5911	0.7489	0.7628	0.7572

For humidity, the models showed a correlation coefficient of 0.93 and an R^2 value of 0.86 when considering all data, indicating high accuracy. In urban areas, the highest R-value was 0.95, with an R^2 value of 0.90, indicating very accurate predictions. Suburban areas achieved R = 0.95 and R^2 = 0.90, similar to urban areas. In rural areas, the R-value was 0.87 and the R^2 value was 0.75, indicating a reasonable level of accuracy, though lower than in urban and suburban areas.

For solar irradiation, the correlation coefficient for all data was 0.91, with an R^2 value of 0.83, reflecting strong model performance. In urban areas, the highest R-value was 0.97, with an R^2 value of 0.93, indicating very high accuracy. Suburban areas achieved R = 0.95 and R^2 = 0.91, showing strong model accuracy. In rural areas, the R-value was 0.87 and the R^2 value was 0.76, which, although lower, still indicated good model performance. For pressure, the models showed a correlation coefficient of 0.81 and an R^2 value of 0.64 when considering all data, indicating moderate accuracy. In urban areas, the highest R-value was 0.98, with an R^2 value of 0.95, reflecting excellent model performance. Suburban areas showed strong performance with R = 0.92 and R^2 = 0.84. In rural areas, the R-value was 0.88 and the R^2 value was 0.76, indicating a reasonable level of accuracy.

The results from the model verification phase reveal that the models perform exceptionally well in urban and suburban areas, with high correlation coefficients and R^2 values across all meteorological parameters. However, the accuracy is slightly lower in rural areas, which may be due to the more variable environmental conditions in these areas and due to the lower number of input data. Overall, the models for temperature, humidity, and solar irradiation exhibit very high accuracy, particularly in urban and suburban areas. The model for pressure, while still accurate, shows moderate performance overall but excellent results in urban settings. These findings underscore the robustness and reliability of the developed models in different geographical contexts, providing valuable insights for agricultural applications and real-time monitoring.

5. Conclusions

This research aimed to solve the challenges caused by climate changes, particularly extreme temperature events and drought, using Internet of Things technologies. Accurate weather predictions enable better irrigation planning and crop management, ultimately improving yield and reducing costs. Our study emphasizes the importance of high-precision weather forecasting models developed using advanced machine-learning techniques and extensive IoT sensor networks. The integration of these technologies provides a valuable tool for farmers, enabling informed decision-making and optimizing agricultural practices to cope with the challenges posed by climate change. To monitor the effects of drought on maize crops in the Republic of Croatia, we analyzed a dataset of 139,965 weather data points collected in the summer of 2022 using 18 commercial sensor nodes operating with the LoRaWAN protocol. The data included measurements of temperature, humidity, solar irradiation, and air pressure. We developed forecasting models intended for future use

in maize crops, taking into account the impact of urbanization on agrometeorological parameters. We divided the data into urban, suburban, and rural segments to fill gaps in the existing literature. Our approach was to analyze the data using 19 different regression modeling techniques, creating four regional models per parameter and four general models applicable to all areas. This comprehensive analysis allowed us to identify the most effective models for each area, improving the accuracy of our agrometeorological forecasts and optimizing maize yields under changing weather conditions. The research integrates ML and A with the IoT for agriculture, providing innovative solutions for predictive analytics in crop production. By focusing on solar radiation in addition to traditional weather parameters and taking geographical differences into account, our models provide a tool to improve agricultural sustainability in times of climate change.

In the verification phase, the models were evaluated using Pearson's correlation coefficient and R^2 in urban, suburban, and rural areas. The results showed that the models performed exceptionally well in both urban and suburban areas, with high correlation coefficients and R^2 values for all parameters. In rural areas, the accuracy was slightly lower, probably due to the more variable environmental conditions and less input data. In general, the models for temperature, humidity, and solar radiation showed very high accuracy, especially in urban and suburban areas. The air pressure model, while accurate, performed moderately well overall but was excellent in urban areas. Our results highlight the effectiveness of using advanced regression models, such as GPR and tree-based models, in analyzing meteorological data for agricultural applications. These models provide effective real-time monitoring and analysis tools and increase the accuracy of agricultural forecasts. The comprehensive database and powerful models developed and open-sourced in this study provide valuable resources for future research and practical applications in smart agriculture.

In future research, the presented database should be expanded by deploying additional sensor nodes across various regions, to enhance the model's efficiency in rural areas. Additionally, further investigation into air pressure estimation is warranted to improve accuracy and reliability considering it has not yet been investigated enough. Thoroughly examining models for predicting weather parameters in micro-locations is essential to ensure localized and precise forecasts. This includes a detailed analysis of model performance across different geographical settings and the integration of advanced machine-learning techniques to refine predictions for temperature, humidity, air pressure, and solar irradiation in specific micro-climates.

Author Contributions: Conceptualization, J.Š. and J.S.; methodology, J.Š. and J.S.; software, J.Š. and J.S.; validation, J.Š.; formal analysis, J.Š.; investigation, J.Š. and J.S.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.Š. and J.S.; writing—review and editing, K.G., D.Ž., and J.S.; visualization, J.Š.; supervision, K.G. and D.Ž.; project administration, K.G. and D.Ž.; funding acquisition, D.Ž. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the project "IoT-field: An Ecosystem of Networked Devices and Services for IoT Solutions Applied in Agriculture" co-financed by the European Union from the European Regional Development Fund within the Operational Program Competitiveness and Cohesion 2014–2020 of the Republic of Croatia.

Data Availability Statement: A dataset of 139,965 weather data points collected in the summer of 2022 using 18 commercial sensor nodes operating with the LoRaWAN protocol can be downloaded at: https://github.com/jspisi88/Agri-Weather-Prediction-ML (accessed on 15 August 2024).

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

References

1. Singh, S.; Kaushik, M.; Gupta, A.; Malviya, A.K. Weather Forecasting Using Machine Learning Techniques. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), Sultanpur, India, 11 March 2019.
2. Maheswari, K.B.; Gomathi, S. A Comprehensive Analysis of Weather Prediction Using Machine Learning. In Proceedings of the Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 5–6 April 2024.
3. Guerra, J.C.V.; Khanam, Z.; Ehsan, S.; Stolkin, R.; McDonald-Maier, K. Weather Classification: A New Multi-Class Dataset, Data Augmentation Approach and Comprehensive Evaluations of Convolutional Neural Networks. In Proceedings of the 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS), Edinburgh, UK, 6–9 August 2018.
4. Wadhwa, S.; Tiwari, R.G. Machine Learning-Based Weather Prediction: A Comparative Study of Regression and Classification Algorithms. In Proceedings of the 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), Bhubaneswar, India, 9–11 June 2023.
5. Agarwal, A.B.; Rajesh, R.; Arul, N. Spatially-Resolved Hyperlocal Weather Prediction and Anomaly Detection Using IoT Sensor Networks and Machine Learning Techniques. In Proceedings of the 2023 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom), Dubai, United Arab Emirates, 7–9 December 2023.
6. Somasundaram, R.S.; Nagamani, K.; Lilly Florence, M.; Swamydoss, D. Estimation and Prediction of Crop Yielding Rate Using Machine Learning Techniques. In Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 25–27 January 2022.
7. Sadhukhan, M.; Dasgupta, S.; Bhattacharya, I. An Intelligent Weather Prediction System Based on IoT. In Proceedings of the 2021 Devices for Integrated Circuit (DevIC), Kalyani, India, 19–20 May 2021.
8. Hossain, M.S.; Mahmood, H. Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast. *IEEE Access* **2020**, *8*, 172524–172533. [CrossRef]
9. Vogfjörd, K.; Jakobsdóttir, S.; Gudmundsson, G.; Roberts, M.; Ágústsson, K.; Arason, T.; Geirsson, H.; Karlsdóttir, S.; Hjaltadóttir, S.; Ólafsdóttir, U.; et al. Forecasting and Monitoring a Subglacial Eruption in Iceland. *Eos Trans. AGU* **2005**, *86*, 245–248. [CrossRef]
10. Siddique, T.; Mahmud, M.S.; Keesee, A.M.; Ngwira, C.M.; Connor, H. A Survey of Uncertainty Quantification in Machine Learning for Space Weather Prediction. *Geosciences* **2022**, *12*, 27. [CrossRef]
11. Kolla, V.R.K. Forecasting the Future: A Deep Learning Approach for Accurate Weather Prediction. *Int. J. IT Eng. (IJITE)* **2018**, *6*, 37–44.
12. Lima, M.A.F.B.; Carvalho, P.C.M.; Fernandez-Ramírez, L.M.; Braga, A.P.S. Improving Solar Forecasting Using Deep Learning and Portfolio Theory Integration. *Energy* **2020**, *195*, 117016. [CrossRef]
13. Zhou, Y. Advances of Machine Learning in Multi-Energy District Communities—Mechanisms, Applications and Perspectives. *Energy AI* **2022**, *10*, 100187. [CrossRef]
14. NOAA National Centers for Environmental Information. Climate Data Online. Available online: https://www.ncdc.noaa.gov/ (accessed on 5 July 2024).
15. European Centre for Medium-Range Weather Forecasts. ERA5 Reanalysis Dataset. Available online: https://www.ecmwf.int/ (accessed on 5 July 2024).
16. Pauzi, A.F.; Hasan, M.Z. Development of IoT Based Weather Reporting System. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *917*, 012032. [CrossRef]
17. Nsabagwa, M.; Byamukama, M.; Kondela, E.; Sansa, O.J. Towards a Robust and Affordable Automatic Weather Station. *Dev. Eng.* **2019**, *4*, 100040. [CrossRef]
18. Mohapatra, D.; Subudhi, B. Development of a Cost Effective IoT-based Weather Monitoring System. *IEEE Consum. Electron. Mag.* **2022**, *11*, 81–86. [CrossRef]
19. National Aeronautics and Space Administration. Goddard Earth Sciences Data and Information Services Center. Available online: https://disc.gsfc.nasa.gov/ (accessed on 5 July 2024).
20. World Meteorological Organization. WMO Public Weather Services. Available online: https://public.wmo.int/en (accessed on 5 July 2024).
21. Meteostat. Historical Weather Data. Available online: https://meteostat.net/ (accessed on 5 July 2024).
22. Kaggle. Datasets for Machine Learning. Available online: https://www.kaggle.com/datasets (accessed on 5 July 2024).
23. Global Historical Climatology Network (GHCN). Daily Data. Available online: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily (accessed on 5 July 2024).
24. OpenWeatherMap. Weather Data API. Available online: https://openweathermap.org/ (accessed on 5 July 2024).
25. Rokach, L.; Maimon, O. Decision Trees. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 165–192.
26. Tercha, W.; Tadjer, S.A.; Chekired, F.; Canale, L. Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems. *Energies* **2024**, *17*, 1124. [CrossRef]
27. Priyadarshi, H.; Singh, K.; Shrivastava, A. Insolation Prediction for Reliable Energy Storage Using Deep Learning Algorithms. In Proceedings of the 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 27–28 August 2021., 27–28 August 2021.

28. Haupt, S.E.; Cowie, J.; Linden, S.; McCandless, T.; Kosovic, B.; Alessandrini, S. Machine Learning for Applied Weather Prediction. In Proceedings of the 2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam, The Netherlands, 29 October–1 November 2018.
29. Scher, S.; Messori, G. Predicting Weather Forecast Uncertainty with Machine Learning. *Q. J. R. Meteorol. Soc.* **2018**, *144*, 2830–2841. [CrossRef]
30. Pisner, D.A.; Schnyer, D.M. Support Vector Machine. In *Machine Learning: Methods and Applications to Brain Disorders*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 101–121.
31. Jun, Z. The Development and Application of Support Vector Machine. *Stata J. Promot. Commun. Stat. Stata* **2020**, *20*, 3–29. [CrossRef]
32. Rasmussen, C.E.; Williams, C.K.I. Gaussian Processes for Machine Learning. In *Advanced Lectures on Machine Learning*; Bousquet, O., Luxburg, U., Rätsch, G., Eds.; The MIT Press: Canberra, Australia, 2006; Volume 3176, pp. 63–71.
33. Hamasuna, Y.; Yokoyama, Y.; Takegawa, K. The Relationship Between Gaussian Process Based C-Regression Models and Kernel C-Regression Models. In Proceedings of the 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Ise, Japan, 29 November–2 December 2022.
34. Javed, A.; Kasi, B.K.; Khan, F.A. Predicting Solar Irradiance Using Machine Learning Techniques. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019.
35. Kirch, W. Pearson's Correlation Coefficient. In *Encyclopedia of Public Health*; Kirch, W., Ed.; Springer: Dresden, Germany, 2008; Volume 1, pp. 1090–1091.
36. Willmott, C.; Matsuura, K. Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.* **2005**, *30*, 79. [CrossRef]
37. Diniz, P.S.R. Fundamentals of Adaptive Filtering. In *Adaptive Filtering*; Diniz, P.S.R., Ed.; The Springer International Series in Engineering and Computer Science; Springer: Boston, MA, USA, 1997; Volume 399, pp. 1–10.
38. Mierzwiak, M.; Kroszczyński, K. Impact of Domain Nesting on High-Resolution Forecasts of Solar Conditions in Central and Eastern Europe. *Energies* **2023**, *16*, 4969. [CrossRef]
39. Wolniak, R.; Skotnicka-Zasadzień, B. Development of Photovoltaic Energy in EU Countries as an Alternative to Fossil Fuels. *Energies* **2022**, *15*, 662. [CrossRef]
40. Verma, G.; Mittal, P.; Farheen, S. Real Time Weather Prediction System Using IoT and Machine Learning. In Proceedings of the 2020 6th International Conference on Signal Processing and Communication (ICSC), Noida, India, 5–7 March 2020.
41. Shahin, M.B.U.; Sarkar, A.; Sabrina, T.; Roy, S. Forecasting Solar Irradiance Using Machine Learning. In Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry, Dhaka, Bangladesh, 19–20 December 2020.
42. METEOHELIX® IoT PRO Dataset. Available online: https://static1.squarespace.com/static/597dc443914e6bed5fd30dcc/t/656d99b8f279294f7ab2db3a/1701681664838/MeteoHelix+IoT+Pro+DataSheet.pdf (accessed on 5 July 2024).
43. LoRaWAN Weather Station (WS100LRW/LW) Dataset. Available online: https://senetco.com/marketplace/ubiq-iot-ws100lrw-weather-station/ (accessed on 5 July 2024).
44. Eleven Parameter Weather Station for LoRaWAN® Dataset. Available online: https://cdn.decentlab.com/download/datasheets/Decentlab-DL-ATM41-datasheet.pdf (accessed on 5 July 2024).
45. Agri-Weather Prediction Dataset and Analysis. Available online: https://github.com/jspisi88/Agri-Weather-Prediction-ML (accessed on 15 August 2024).
46. Kumar, M.; Namrata, K.; Kumar, N.; Saini, G. Solar Irradiance Prediction Using an Optimized Data Driven Machine Learning Models. *J. Grid Comput.* **2023**, *21*, 28. [CrossRef]
47. Budiyanto, M.A.; Lubis, M.H. Comparison Result of Hourly Solar Radiation under the Clear Sky Condition Based on Solar Radiation Model and Measured Data Experiment. In Proceedings of the 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE), Yogyakarta, Indonesia, 13–14 October 2020.
48. Ozbek, A.; Ünal, Ş.; Bilgili, M. Daily average relative humidity forecasting with LSTM neural network and ANFIS approaches. *Theor. Appl. Climatol.* **2022**, *150*, 697–714. [CrossRef]
49. Cooper, F.; Chantry, M.; Düben, P.; Bechtold, P.; Sandu, I. Statistical Modeling of 2-m Temperature and 10-m Wind Speed Forecast Errors. *Mon. Weather Rev.* **2023**, *151*, 897–911.
50. Zhang, Z.; Dong, X.; Xue, Y.; Jiang, M.; Cao, H. Temperature Prediction for Ultra-Low Temperature Measurement System Based on Improved Extreme Learning Machine Algorithm. In Proceedings of the 2023 3rd International Conference on Electronic Information Engineering and Computer Communication (EIECC), Wuhan, China, 8–10 December 2023.

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.