agriculture

Article

Smart Weather Data Management Based on Artificial Intelligence and Big Data Analytics for Precision Agriculture

Chouaib El Hachimi [1,*], Salwa Belaqziz [1,2], Saïd Khabba [1,3], Badreddine Sebbar [1,4], Driss Dhiba [5] and Abdelghani Chehbouni [1,5]

1 Center for Remote Sensing Applications (CRSA), Mohammed VI Polytechnic University (UM6P), Ben Guerir 43150, Morocco
2 LabSIV Laboratory, Department of Computer Science, Faculty of Science, UIZ University, Agadir 80000, Morocco
3 LMFE, Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakesh 40000, Morocco
4 Centre d'Etudes Spatiales de la Biosphère (CESBIO), Université de Toulouse, 31400 Toulouse, France
5 International Water Research Institute (IWRI), Mohammed VI Polytechnic University (UM6P), Ben Guerir 43150, Morocco
* Correspondence: chouaib.elhachimi@um6p.ma

Abstract: Smart management of weather data is an essential step toward implementing sustainability and precision in agriculture. It represents an important input for numerous tasks, such as crop growth, development, yield, and irrigation scheduling, to name a few. Advances in technology allow collecting this weather data from heterogeneous sources with high temporal resolution and at low cost. Generating and using these data in their raw form makes no sense, and therefore implementing adequate infrastructure and tools is necessary. For that purpose, this paper presents a smart weather data management system evaluated using data from a meteorological station installed in our study area covering the period from 2013 to 2020 at a half-hourly scale. The proposed system makes use of state-of-the-art statistical methods, machine learning, and deep learning models to derive actionable insights from these raw data. The general architecture is made up of four layers: data acquisition, data storage, data processing, and application layers. The data sources include real-time sensors, IoT devices, reanalysis data, and raw files. The data are then checked for errors and missing values using a proposed method based on ERA5-Land reanalysis data and deep learning. The resulting coefficient of determination (R^2) and Root Mean Squared Error (RMSE) for this method were 0.96 and 0.04, respectively, for the scaled air temperature estimate. The MongoDB NoSQL database is used for storage thanks to its ability to deal with real-world big data. The system offers various services such as (i) weather time series forecasts, (ii) visualization and analysis of meteorological data, and (iii) the use of machine learning to estimate the reference evapotranspiration (ET_0) needed for efficient irrigation. To this, the platform uses the XGBoost model to achieve the precision of the Penman–Monteith method while using a limited number of meteorological variables (air temperature and global solar radiation). Results for this approach give R^2 = 0.97 and RMSE = 0.07. This system represents the first incremental step toward implementing smart and sustainable agriculture in Morocco.

Keywords: artificial intelligence; big data analytics; smart agriculture; evapotranspiration; ERA5-Land; time series forecasting; anomaly detection; MongoDB

Citation: Hachimi, C.E.; Belaqziz, S.; Khabba, S.; Sebbar, B.; Dhiba, D.; Chehbouni, A. Smart Weather Data Management Based on Artificial Intelligence and Big Data Analytics for Precision Agriculture. *Agriculture* **2023**, *13*, 95. https://doi.org/10.3390/agriculture13010095

Academic Editor: Bin Gao

Received: 6 December 2022
Revised: 22 December 2022
Accepted: 23 December 2022
Published: 29 December 2022

1. Introduction

Advances in technology and industry have helped humanity to increase life quality and expectancy. This includes delegating laborious processes traditionally performed by hand to machines that can perform these tasks more efficiently. However, it has brought with it several problems as well [1] relating to the overexploitation and nonrational use of Earth's natural resources, which in turn causes disruption of the natural balance expressed

by climate change effects, such as global warming and severe climate events (droughts, flooding, storms, hurricanes, etc.). In response to this alarming situation, humanity is asked more than ever to rethink and reconsider its way of dealing with the environment, especially in the context of a world population that grows at a higher rate [2]. This puts a lot of pressure on our food systems to meet the increasing demand and feed the planet. Agriculture is the sector that must be addressed, and agricultural management practices need to be optimized and become more efficient and sustainable to address these challenges. With water resources as an example, agriculture comes in as the largest consumer of this resource, with about an average of 70% in use around the world [3]. In this sector, water resources are used mainly in irrigation activities that still follow unsustainable methods such as gravity irrigation, the most widely used type of irrigation globally [4,5]. Additionally, even when using modern methods of irrigation, such as drip irrigation, they can be inefficient in the absence of good management [6]. The crucial step to achieving an efficient irrigation system, whatever the type of irrigation system used, is to know the proper amount of water to supply and the time to apply it (when and how much). To this end, we need to monitor either the soil, the crop, or the weather. Monitoring the weather enables the estimation of evapotranspiration (ET), which is the sum of transpiration from plants and soil surface evaporation. Accurately estimating this parameter allows compensation of the lost water quantity to the atmosphere. There are two methods of estimating the evapotranspiration of a crop. One uses a single crop coefficient (K_c) that incorporates all physiological and physical variations between the crops and the second where the K_c is split into two separate coefficients: K_{cb} for crop transpiration and K_e for soil evaporation. The K_c coefficient, then, is multiplied by the reference evapotranspiration (ET_0), which reflects the rate of evapotranspiration for a specific crop (grass). It is obtained by monitoring various meteorological parameters (air temperature, solar radiation, air relative humidity, wind speed, etc.). Therefore, weather monitoring is a key step toward implementing efficient irrigation systems and ensuring sustainable agriculture.

Today, we are able to collect weather data with high spatial and temporal resolution thanks to advances in science and technology, including advances in remote sensing such as the availability of Unmanned Aerial Vehicle (UAV) [7] equipped with cutting-edge sensors with affordable prices, satellite imagery with different spatio-temporal resolutions, open-access reanalysis data such as European ReAnalysis data [8], MERRA [9], the JRA-55 [10] or NCEP–DOE AMIP-II Reanalysis data [11]. Additionally, advances in the Internet of Things (IoT) field have enabled cost-effective sensor data acquisition [12–16]. This huge amount of data raised issues regarding the efficiency, complexity, interfaces, dynamics, robustness, and interaction between these new types of peer-to-peer connected systems that need to be re-examined on a large scale, as discussed in [17]. This unprecedented amount of generated data is also contributing to what is known as big data, which needs adequate infrastructure to be stored and gain insights for assisting agricultural decision making.

This work aligns with this objective and is intended to provide intelligent ways to deal with these abundant data by applying the concepts of big data analytics and leveraging the potential applications of artificial intelligence in agriculture to help farmers minimize their risks, or at least make them more manageable, which represents a step towards the development of smart and precision agriculture.

2. State of the Art

Several researchers have tried to leverage the challenges and opportunities of the big data wave in the agricultural field and how it can be the driver of sustainability and precision in this sector by gaining insights from massive volumes of data that can be used to assist decision making.

As surveyed in [18], big data analytics is leading to advances in various industries, but it has not yet been widely applied in agriculture. The work presents a range of suggested solutions, tools, algorithms, and data, including how they were used and their impact upon the sector as a whole. It also emphasizes the enormous potential of big data analytics in

agriculture for smarter farming, demonstrating how the accessibility of big data analytics tools, methods, and software, as well as the growing openness of heterogeneous data sources (the "open data wave"), will encourage more academic research, public sector initiatives, and business ventures in the agricultural sector. The paper concludes that the adoption of big data practices in agriculture is still challenging and faces several obstacles when applied to real-world applications.

In the paper [19], the authors developed a system composed of three components: (i) hardware to capture crop data, (ii) a web application for crop data and field information manipulation, and (iii) a mobile application to control irrigation via mobiles. The objective of this work is to analyze the suitability of crops in terms of air temperature, air relative humidity, and soil moisture to optimize future plans and strategies.

The authors of the work [20] propose a solution to be adopted in India to help farmers face the unpredictable nature and variability of climate and weather circumstances. They built a system for managing local weather stations in real time that would keep farmers well-informed about the current weather conditions in advance, allowing them to make the right decisions at the right time and prevent crop loss. High-speed internet infrastructure available even in rural areas is the main motivation for conducting this work. It facilitates the communication of collected data to remote servers. Once received and analyzed, the information derived from these data gives farmers a way to automate their agricultural management practices (irrigation, fertilization, and harvesting) by triggering the right action at the right time.

Article [21] also presented a smart weather station management system intended to be used in agriculture and to manage meteorological stations. It is based on Internet of Things technology (IoT) to minimize costs. The connected sensors measure air temperature, humidity, light intensity, air pressure, and wind speed. It then sends the collected data to the server part of the system through the Global System for Mobile Communications (GSM) module. The application layer is powered by the ThingSpeak platform, which offers standard services to the proposed system such as data visualization and data analysis.

In regards to dealing with the challenges related to handling huge amounts of data generated by sensors, the paper [22] proposed a standard architecture for a data infrastructure platform called WALLeSMART, which is a cloud-based solution that provides a general architecture to handle the difficulties of gathering, processing, storing, and visualizing extremely large amounts of data in batch and real-time modes. An initial prototype has been developed and tested at various farms in the Wallonia region of Belgium, showing prominent results. This proposed system can be used as the basis for developing customized smart agricultural services to meet our needs.

Our contributions not only propose an architecture containing some of the standard pipelines used in the literature to build data platforms, namely data acquisition, storing, visualizing, and analysis, but a complete system with the aim of going from data to decision making. The proposed system provides services such as weather time series forecasting, missing values handling using a multisource approach (reanalysis and situ data), and estimating important parameters needed in the day-to-day life of farmers, such as the evapotranspiration (ET).

3. Study Area

The study area is situated 40 km east of Marrakesh in the semiarid Haouz plain in the heart of Morocco (Figure 1). About 2800 ha of this area is irrigated, and it is nearly flat. Cereal crops such as wheat and barley are the main dominant crops. The region's climate is typically Mediterranean semiarid, with an average annual rainfall of about 250 mm [23,24], temperatures that range from hot in the summer (38 °C in July) to cool in the winter (5 °C in February), significant daily and monthly variations that are concentrated primarily from autumn to spring, and an average annual ET_0 of 1600 mm [25].

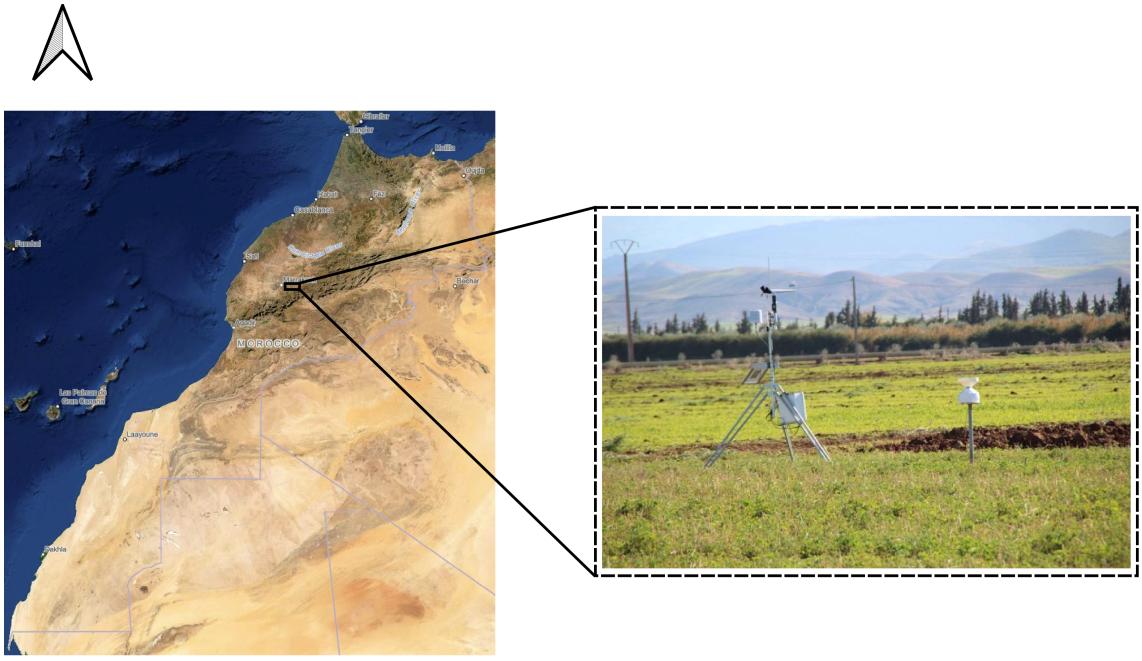

Figure 1. R3 district study area in Morocco and a Photo of the meteorological station installed.

4. System Architecture

The proposed system is designed to enable the smart management of weather data, which represents the key to implementing smart agriculture. By monitoring and analyzing the weather effectively, we can optimize various agricultural management practices such as irrigation scheduling and choosing the appropriate crop to sow [26].

The design of the platform follows a service-oriented architecture (Figure 2) to offer services that address each of the four categories of big data analytics. As part of a descriptive data analysis that tries to understand what happened in the past, a scenario would be: rainfall declined, and the frequency changed over the last decade. The answer to why this happened, in turn, is very important, and it takes us to the diagnostic data analysis, where the focus is to identify anomalies in data to explain the reasons behind events, such as linking this event with long-term shifts in temperatures and weather patterns observed in weather evolution charts. The third type covered by the platform is predictive data analysis. It looks beyond the present and tries to predict the future using statistical methods and machine learning algorithms that learn from historical data in an iterative approach, trying to identify the optimal way to predict the future. One such service is weather forecasting. The output of the forecasting service can be used to support decision making about what actions to take that aim to prevent severe events from occurring in the future. This is carried out through the last type, which is prescriptive data analysis.

The platform can be decomposed into four main layers: the data acquisition layer, the data storage layer, the processing layer, and the application layer.

4.1. Data Acquisition Layer

Heterogeneous data from a variety of sources, including meteorological station data, IoT weather sensors, reanalysis data, third-party meteorological services, and raw files (CSV, Excel, etc.), are used as the input for this layer. The concept of big data is introduced to the field by the volume, velocity, and variety that characterize these data. Missing values are handled in this layer using the method developed in Section 4.4.1 prior to being stored in the NoSQL MongoDB database.

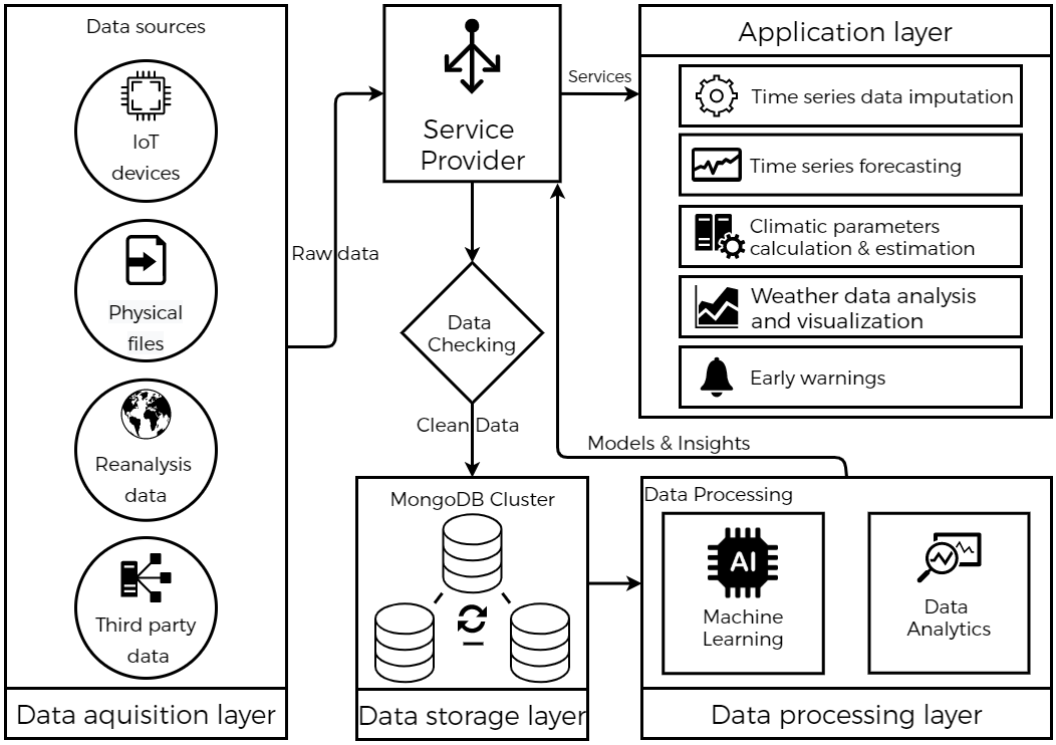

Figure 2. General architecture of the platform.

4.1.1. Weather Station Data

The weather dataset used in this study was collected from a meteorological station installed in the study area (Figure 1) at the half-hour scale from 2013 to 2020. The used tower is equipped with different sensors [27,28] to measure:

- Incoming solar radiation using (Kipp and Zonen CM5 Pyranometer, Delft, The Netherlands).
- Air temperature in Kelvin, relative humidity (R3_Hr, as a fraction between 0 and 1) and vapor pressure by using (HMP45C, Vaisala, Helsinki, Finland).
- Wind speed using (A100R Anemometer, R.M. Young Company, Traverse City, MI, USA).
- Rainfall using (FSS500 Tipping Bucket Automatic Rain Gauge, Campbell Scientific Inc., Logan, UT, USA).

Next, records are stored in data loggers before being collected manually by agents or sent to a centralized server via a cellular connection.

A full description of these data is shown in Table 1, which also includes statistics for missing values.

Table 1. Meteorological station data description.

Variables	Description	Unit	Missing Values
R3_Dv	Wind direction	Degree	2626
R3_Hr	Relative air humidity	No unit	2626
R3_Rg	Global solar radiation	W m^{-2}	5169
R3_Tair	Air temperature	°C	2631
R3_Vv	Wind speed	m s^{-1}	2626
R3_P30m	Rainfall	mm	2626

4.1.2. ERA5-Land Reanalysis Data

Advances in measurement technologies have enabled us to use various observation methods to monitor the Earth's weather, including weather stations, weather balloons, and satellite imagery, to name a few. However the distribution of these observation methods

does not cover the entire globe, they may have overlapping footprints between covered areas, and fewer of them were available in the past, which makes it challenging to conduct studies of past years. To deal with this, climate reanalysis emerges as a new way of trying to deliver a complete picture of the past and of the entire globe by combining the laws of physics, modern weather models, and available weather sources. Such data, if accurate, are crucial and will certainly assist decision making in several domains such as smart cities, smart management of renewable energy stations, sustainable and climate-smart agriculture [29,30], climate change assessments, hydrology [31], and much more.

In our study, we used the fifth generation of European ReAnalysis (ERA5-Land) [32] available to be downloaded for free through the Climate Data Store (CDS) web platform [33]. ERA5-Land is the successor to ERA5 [8], which in turn is the successor to ERA-Interim [34]. This new product covers the period from 1950 to the present. ERA5-Land has the benefit over the predecessors of its high horizontal resolution (9 km against 31 km for ERA5 and 80 km for ERA-Interim). These strengths were achieved thanks to the integration of the ECMWF land surface model forced by the ERA5 climate reanalysis with corrected elevation for the thermodynamic near-surface state and then applied to Numerical Weather Prediction (NWP) models. The data used in this study concern the two pixels that cover our study area (Figure 3). We downloaded ERA5-Land weather data from 2013 to 2020. Since the cloud service Climate Data Store API (cdsapi) has a limit of 100,000 records per request, we downloaded each year separately and combined them all at the end. The downloaded data were then converted to a pandas dataframe data structure using the Python language. We also developed a function called "get_era5_land_grib_as_dataframe", available as a part of the GIS class of the public library Data Science Toolkit (DST) [35] that accepts an ERA5-Land grip file as a parameter and returns a pandas dataframe, which is the most commonly used data structure in the data science field. The full description of the downloaded parameters is described in Table 2.

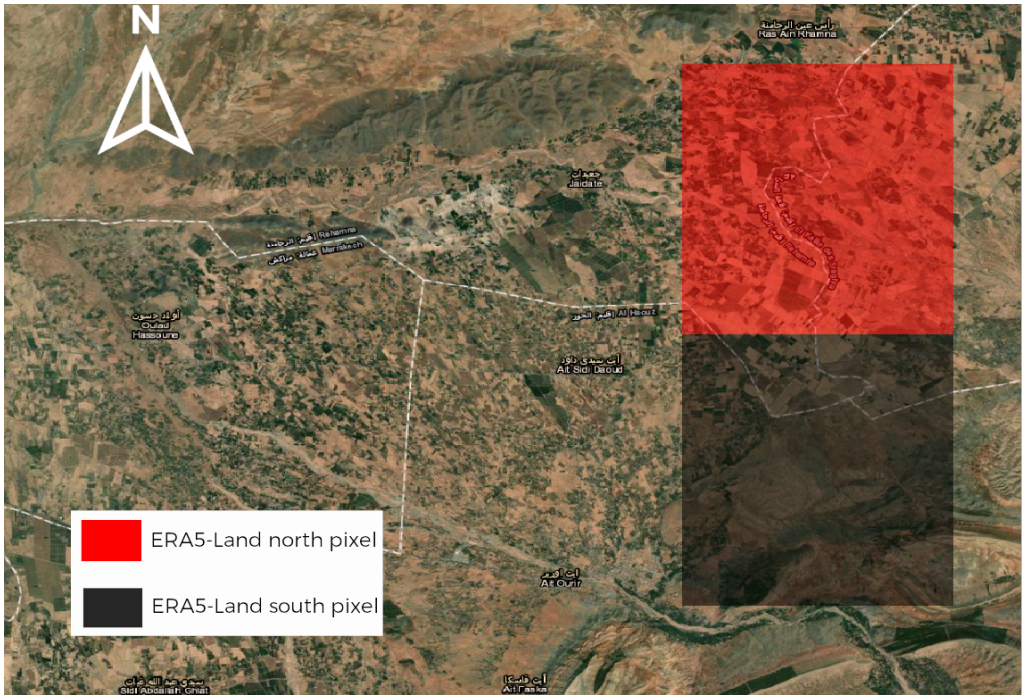

Figure 3. The projection of ERA5-Land pixels over the study area.

Table 2. ERA5-Land downloaded data description.

Variables	Name	Description	Unit
Air temperature	*t2m*	Temperature of air at 2 m above the surface of land.	K
Surface solar radiation downwards	*ssrd*	Amount of solar radiation reaching the surface of Earth. It comprises both direct and diffuse solar radiation.	J m^{-2}
Dewpoint temperature	*d2m*	The temperature to which the air, at 2 m above the surface of the Earth, would have to be cooled for saturation to occur.	K

4.2. Data Storage Layer

In our system, most of the collected data are time series (meteorological data, reanalysis data, satellite data, etc.). This type is characterized mainly by variety (multisource), volume, and velocity (each half-hour). To deal with this, the system uses MongoDB, a big-data-driven database for storing and retrieving meteorological data. The choice of using the MongoDB database comes after several strengths it presents and its suitability for our use case, including that it was designed to replace or enhance the classic Relational Database Management Systems (RDBMS), providing it with a variety of additional characteristics such as scalability being schema-less. It is also powerful at handling large amounts of real-time data and efficiently handling memory, as it is written using the C++ programming language. Not to mention the geospatial indexing feature, which makes it perfect for real-time geospatial data collection and analysis. Figure 4 shows the Entity Relationship Diagram of the weather data subcomponent.

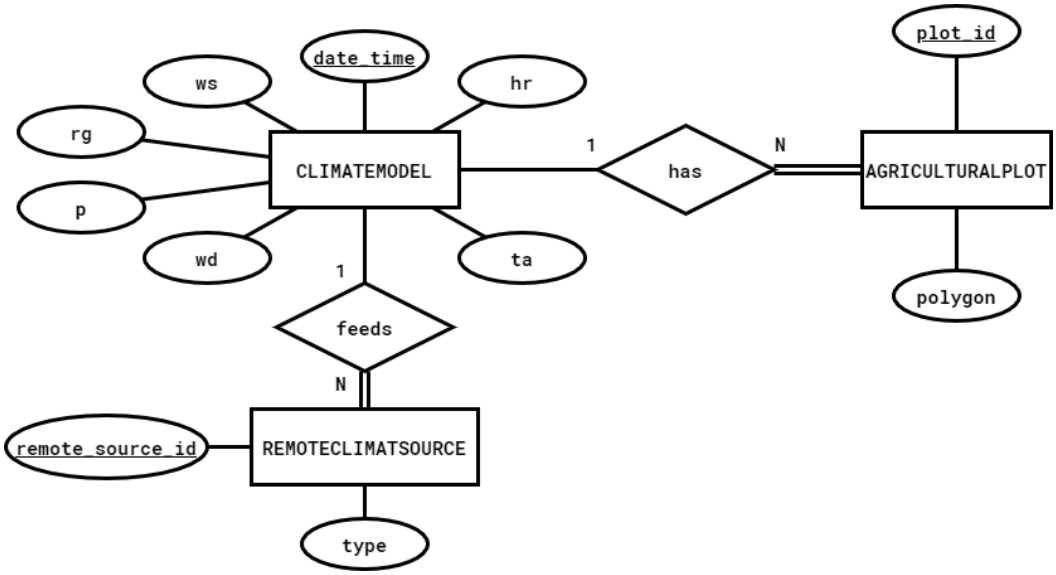

Figure 4. The Entity Relationship Diagram used in the climate database design; ta: air temperature (R3_Tair), rg: global solar radiation (R3_Rg), hr: air relative humidity (R3_Hr), p: rainfall (R3_P), ws: wind speed (R3_Vv), wd: wind direction (R3_Dv).

4.3. Data Processing Layer

The data processing layer takes the data from the data storage layer as input and applies the statistical, machine learning, and deep learning models to gain insights from the data and turn that into services (Figure 2).

4.3.1. Statistical Models

Statistical models are used in this platform for forecasting purposes. Initially, the model Facebook Prophet [36] was used to conduct long-term weather time series forecasting, since it was tested on the same data in previous work.

4.3.2. Machine Learning Models

The system also makes use of machine learning models, given the fact that they can perform well on small datasets, for example, the XGBoost [37] model for reference evapotranspiration estimation based on stored metrological data [38].

4.3.3. Deep Learning Models

Deep learning models or neural networks have gained success in solving complex tasks that were previously human-specific and have required some level of human intelligence to be solved in different fields. They derive their power by trying to mimic the way the human brain works. They are composed of neurons able to process huge amounts of data in order to map the output from a set of inputs using internal mathematical operations. These neurons are organized into groups called layers and, during the propagation of signals between layers in two senses (Forward and backpropagation), the deep learning network learns to perform tasks. In our case, this is a regression task, where the input is ERA5-Land reanalysis data and the output is meteorological station data.

4.4. Application Layer

The application layer contains multiple services related to weather times series.

4.4.1. Time Series Data Imputation Service

It is common in real-world meteorological data to have missing values for various reasons. Missing values can be due to a network error or due to technical issues with certain measurement sensors, etc. These missing data can affect the performance of any type of model (machine learning, numerical, physical, etc.). As such, they need to be identified and handled efficiently during the exploratory data analysis (EDA) and preprocessing stages.

There are several techniques for dealing with missing data depending on the use case:

- Deletion: Deleting rows or columns with missing values will remove this unwanted type of data from our dataset, but it may drastically reduce the size of the dataset, especially in the context of data scarcity.
- Imputation in time series data: In the case of a time series with a trend and seasonality, missing data can be replaced using seasonal adjustment, such as using the data from the same period of the previous year, which is the case for most weather data. However, this method may not be as efficient due to changes in weather patterns around the world. In contrast, if the time series do not present a trend or a seasonality, it can be treated the same way as imputation for a normal dataset.
- Imputation in normal datasets: Replacing it using statistical measures of central tendencies such as the mean, median, or mode of a given window of data that require some assumptions about the distribution type of the data to be efficient.

This service proposes an approach based on reanalysis data and artificial intelligence to build models that can learn rules to map ERA5-Land reanalysis data to station meteorological data, which is also useful for the surrounding regions of our study area (Section 3), since we are in relatively homogeneous areas in terms of elevation and climate. We use the ERA5-Land data presented in Section 4.1.2 and two different architectures of deep learning models. The steps to implement the method workflow are shown in Figure 5.

a. Exploratory data analysis

Before implementing the proposed machine learning approach, exploratory data analysis (EDA) was the first exercise we conducted. It enabled us to understand our collected data, as well as to create hypotheses for further analysis and investigation. In this step, we made no underlying assumptions about the variables, and we were guided only by the observed data. We first calculated the correlation matrix (Figure 6). This matrix allowed us to choose potential estimators for each target variable. Table 3 shows in the second column the potential estimators based on correlation coefficients ($|r| > 0.5$ means

high relationships between variables). In the third column are the estimators that are used in the literature.

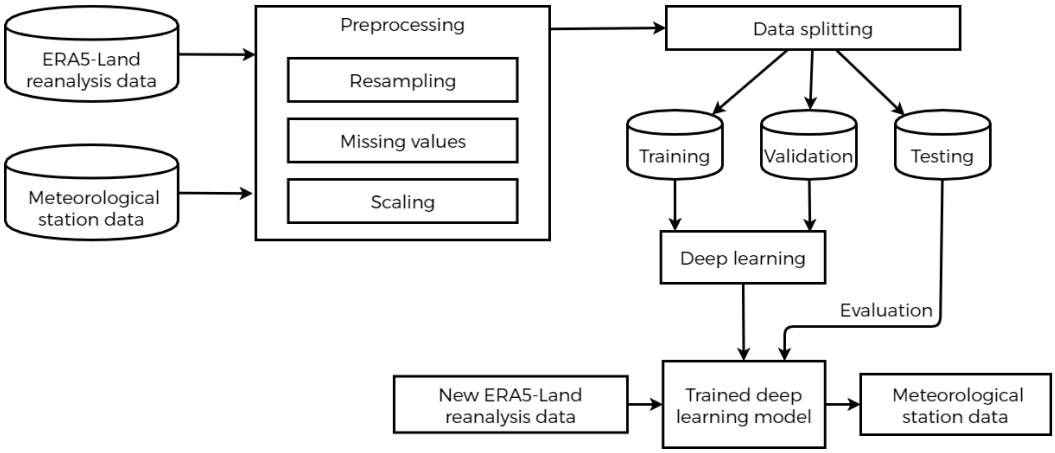

Figure 5. The flowchart of the deep learning approach.

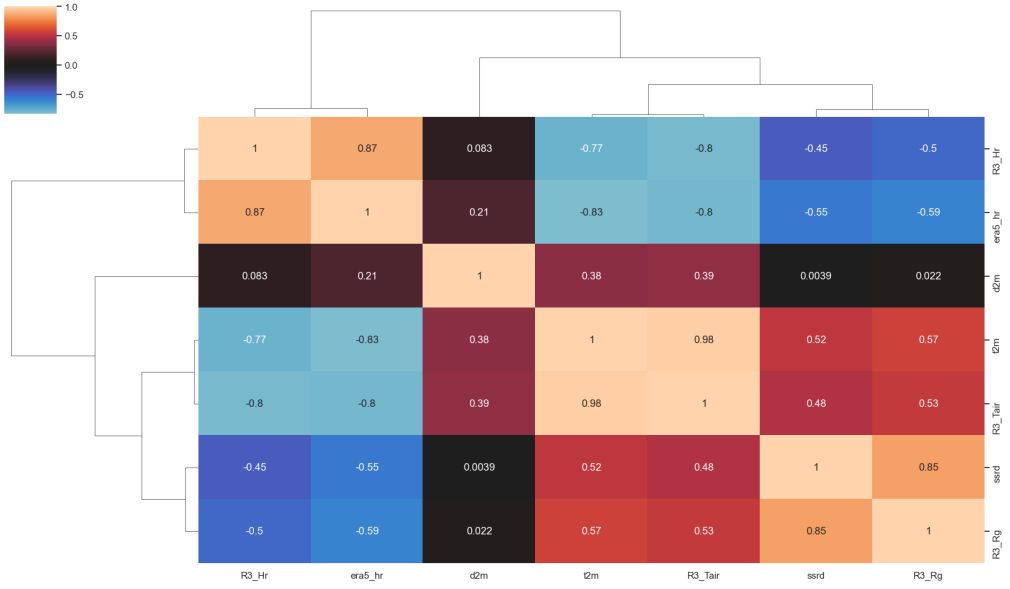

Figure 6. The correlation heatmap and hierarchical clustering of the station and ERA5-Land parameters.

Table 3. Potential estimators of meteorological parameters.

Station Parameter	Correlation Based Potential Estimators	Ground-Truth-Based Potential Estimators
Air temperature (*R3_Tair*)	*t2m*, *R3_Hr* and *R3_Rg*	*t2m*
Global solar radiation (*R3_Rg*)	*ssrd*, *t2m* and *R3_Tair*	*ssrd*
Air relative humidity (*R3_Hr*)	*R3_Tair* and *t2m*	*R3_Tair*, *t2m* and *d2m*

The variable *era5_hr* in the matrix is calculated using the rule of thumb (Equation (1)), which uses both air temperature (*t2m*) and dewpoint temperature (*d2m*) to estimate air relative humidity efficiently for moist air (relative humidity above 50 percent) [39]. According to the matrix, *R3_Hr* has a high negative correlation with *t2m* but a very weak correlation with *d2m*. Despite this, using the latter (*d2m*) combined with *t2m* gives a correlation of 0.87 for *era5_hr* instead of −0.77, if we take only *t2m* into consideration. This motivated us to take both variables as input to neural networks for the estimation of *R3_Hr*. This is also a confirmation for other rule-based (physics-based) approximations that use the

same variables for air relative humidity estimation derived from the Magnus formula [40] (Equation (2)).

$$era5\_hr = 100 - 5(t2m - d2m) \quad (1)$$

$$d2m = \frac{\lambda(\ln(\frac{era5\_hr}{100})) + \frac{\beta.t2m}{\lambda+t2m}}{\beta - (\ln(\frac{era5\_hr}{100}) + \frac{\beta.t2m}{\lambda+t2m})} \quad (2)$$

Equation (2) is valid for air temperatures ranging from −45 to 60 degrees Celsius, the Magnus parameters are, in this case, β = 17.62 and λ = 243.12 degrees Celsius.

For global solar radiation ($R3\_Rg$), the study uses the surface solar radiation downwards (ssrd) as estimator.

Since the objective is to predict meteorological station data based only on ERA5-Land data, we evaluated the performance of two deep learning models, namely a Feed Forward Neural Network (FFNN) and a Long Short-Term Memory (LSTM), to predict $R3\_Tair$ given $t2m$, $R3\_Rg$ given the ssrd, and $R3\_Hr$ given $t2m$ and $d2m$.

b. Feed Forward Neural Network (FFNN):

In this type of neural network, the input signals (ERA5-Land data) are fed into the input layer composed of 100 neurons. In each neuron, the data are processed, taking the weighted sum of inputs plus a bias and then applying an activation function (Figure 7), before forwarding the output to the next layer. The activation function introduces nonlinearity to the output. The choice of this function has a real impact on the training process and performance of models and must be chosen according to the problem at hand. For example, sigmoid and hyperbolic tangent activation functions (Equations (3) and (4)) can be used to capture nonlinearity that may exist between inputs and outputs.

For our case, and given the fact that moving from ERA5-Land reanalysis data to meteorological station data is the inverse of inference in statistics, that is, estimating the mean of an individual (station data) given the mean of the population (ERA5-Land pixels), the error with this assumption is supposed to be linear (polynomial of degree one); that said, we used the Rectified Linear Unit function (ReLU) (Equation (5)) as the activation function in our neural network layers, which is true for both proposed architectures (FFNN and LSTM).

$$y = f(x) = \frac{1}{1 + e^x} \quad (3)$$

$$y = f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4)$$

$$y = f(x) = max(0, x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases} \quad (5)$$

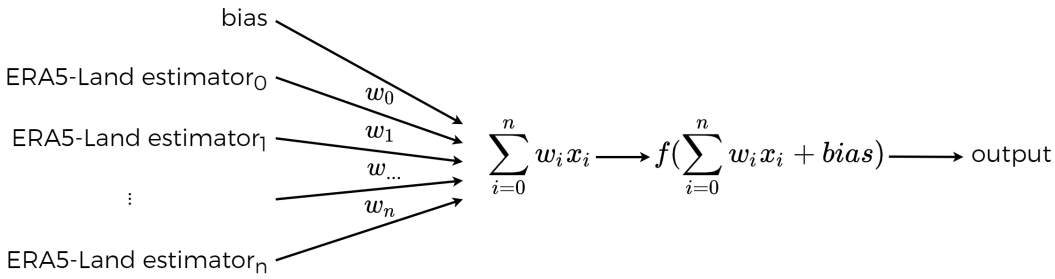

Figure 7. A single neuron model.

c. Long Short-Term Memory (LSTM):

The choice behind using this neural network as a comparison is due to its ability to deal with data that have long-term dependency, which is the case for climate data. LSTM belongs to the Recurrent Neural Networks family (RNN), and therefore also has an internal recurrence, that is, during the learning process, a signal is fed back to a neuron or layer

that has already received and processed it (Figure 8), as well as its ability to remember data through gated cells, which are a kind of memory that accept values in the interval [0, 1] and are used to decide when the flow of a signal passes through the corresponding neuron. LSTM was first developed to resolve the limitations of the vanishing [41] and exploding gradient problems that may occur during the training phase. This problem stops the learning of the neural network because the updates to the various weights become very small.

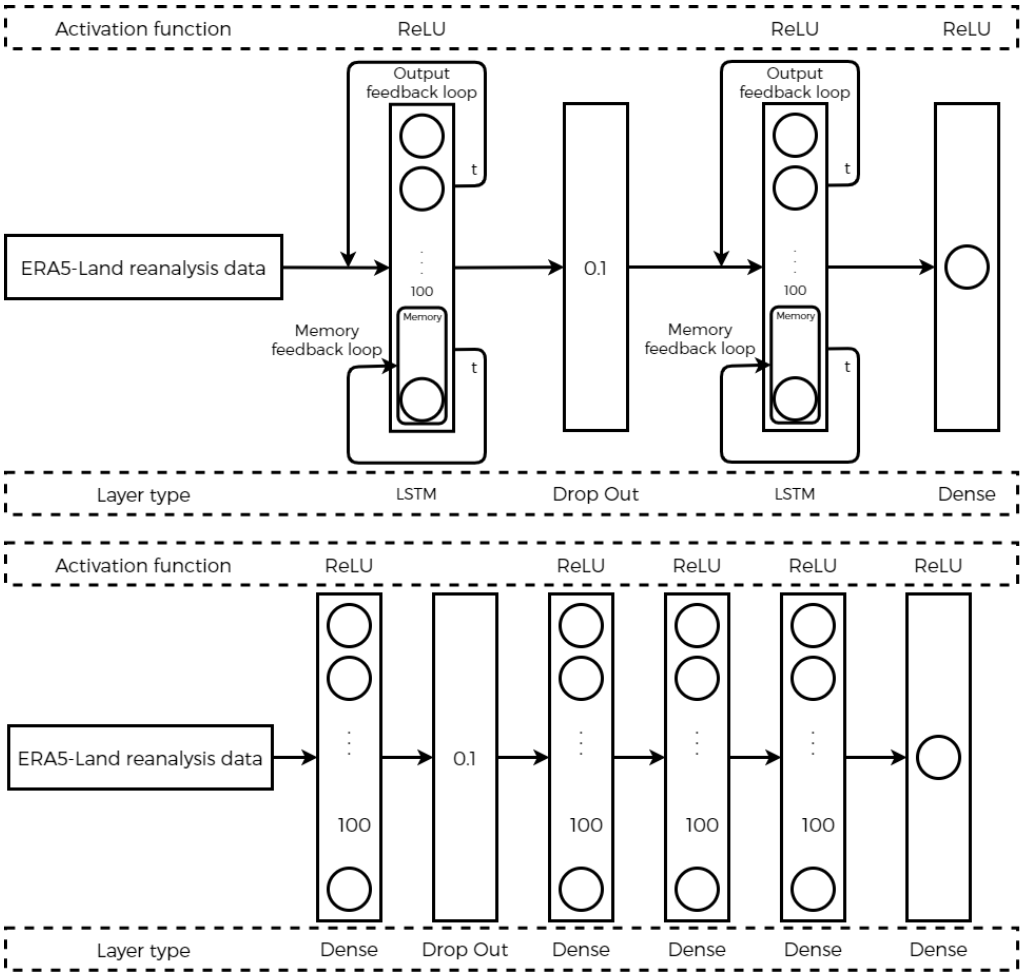

Figure 8. Architectures of FFNN and LSTM used in the approach.

These two architectures (Figure 8) are trained using the Mean Squared Error (MSE) as a loss or cost function that enables calculating the error of a network at the end of a forward pass. To optimize the network weights, we used the adaptive moment estimation optimization algorithm (Adam), which is characterized by fast convergence to the optimal solution and combines the strengths of other optimization algorithms such as Stochastic Gradient Descent (SGD) and RMSProp during the training phase. Like most other optimization algorithms, Adam uses the partial derivative during backward propagation to calculate the error function with respect to each weight within the network (Equation (6)). The Adam algorithm then updates the network weights for a minimized loss or cost function using rules in Equations (7), (8) and (9), respectively.

$$grad = \frac{\partial J}{\partial \theta} \tag{6}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) grad \tag{7}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) grad \tag{8}$$

$$\theta = \theta - \alpha \frac{m_t}{\sqrt{v_t + \epsilon}} \tag{9}$$

d. Data normalization

Data normalization is performed as part of the data preprocessing step and is the process of bringing data to a similar scale. The process is also known as feature scaling. In some cases, such as for statistical machine learning models, it may not be beneficial, but for deep learning models, it is proven to help models to perform better [42,43] in terms of faster convergence, reduced training time, and improved stability (preventing models from oscillating or divergence).

There are multiple methods for data normalization:

- Min–max standardization: Min–max scales the feature values between [0, 1], with 0 being the feature's minimum value and 1 being its maximum value, while maintaining the original distribution (Equation (10)).

$$x_{new} = \frac{x - min(x)}{max(x) - min(x)} \tag{10}$$

- Decimal scaling: This form of scaling is used where values of different decimal ranges are present. For example, two features with different bounds can be brought to a similar scale using decimal scaling (Equation (11))

$$x_{new} = \frac{x}{10^n} \tag{11}$$

  Such that n is an integer representing the order of the scalings.

- Z-score: This transformation scales the value toward a normal distribution with a zero mean and unit variance using the z-score formula (Equation (12)).

$$x_{new} = \frac{x - \mu}{\sigma} \tag{12}$$

  such that μ is the mean and σ is the standard deviation of the features' distribution. This method is very efficient for datasets with a Gaussian distribution.

In our case, we applied the min–max method (Equation (10)). Next, we initialize the hyperparameters of the two proposed architectures (Table 4). These parameters are not updated during the learning phase.

Table 4. Hyperparameters used during the training of FFNN and LSTM models.

Hyperparameter or Layer	FFNN	LSTM
Epochs	20	20
Learning rate	0.0001	0.0001
Batch size	64	64
Layer 1	100 neurons	100 LSTM unit
Layer 2	0.1 for dropout probability	0.1 for dropout probability
Layer 3	100 neurons	100 LSTM unit
Layer 4	100 neurons	1 neuron
Layer 5	100 neurons	—
Layer 6	1 neuron	—

e. Dataset splitting

Before training begins, we split our dataset into three sets: training, validation, and test sets. The validation set is used to assess the performance of the model during each epoch of the training phase. Next, the test set is used to evaluate the final trained model. We used 80–20% splitting for the training–test sets, respectively, and took 20% of the 80% for the validation set.

f. Evaluation Metrics

To evaluate the trained deep learning models' performance on the test dataset, we employed the most commonly used metrics for regression tasks (Equations (13)–(16)):

- Training time: The time it takes for the model to complete 20 epochs.
- R2 score or R^2: The coefficient of determination informs about how well the unknown samples will be predicted by our model. It ranges between 0 and 1, but it can be negative as well (Equation (13)).

$$R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y}_i)^2} \tag{13}$$

- The Pearson correlation coefficient (R): It measures the linear relationship between two normal distributed variables (Equation (14)).

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{14}$$

- Root Mean Squared Error (RMSE): The average of the squares of the errors between real and predicted values by the model (Equation (15)).

$$RMSE = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{y}_i)^2} \tag{15}$$

- Mean Absolute Error (MAE): This is the average of absolute errors between real and predicted values (Equation (16)).

$$MAE = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \tag{16}$$

Given the fact that we applied the scaling in our data using the formula in Equation (10), all metrics (from Equations (13) to (16)) are unitless.

4.4.2. Forecasting Service

The forecasting service helps make projections of the future state of the atmosphere (air temperature, air relative humidity, global solar radiation, etc.) by performing a univariate time series forecasting task. The Facebook Prophet model is used according on the performance it has provided when trained and evaluated using the same meteorological dataset to perform long-term weather forecasting tasks.

4.4.3. Climatic Parameters Calculation and Estimation Service

One of the most important agricultural practices in the day-to-day life of a farmer is irrigation. To achieve efficiency, we need to accurately estimate crops' water needs at each phase of crop growth throughout the agricultural season. This can be performed through several methods, among others is the estimation of the evapotranspiration (ET). It indicates the amount of water loss caused by transpiration from the crop and soil surface evaporation. We can obtain the ET of a crop (ET_c) by multiplying the reference evapotranspiration (ET_0) and crop coefficient K_c, which holds all the physical and physiological differences of a given crop. This service proposes to estimate the ET_0 using machine learning, namely the XGBoost model constrained by the physical model FAO Penman–Monteith. The proposed approach follows the steps presented in Figure 9. First, we resampled the air temperature, global solar radiation, air relative humidity, and wind speed to the daily average, and precipitation to the cumulative daily values, and then the missing values were deleted.

To select the most important contributor variables to the ET_0 estimate, the method uses a random-forest-based technique that ranks the importance of features based on their

occurrences in nodes across all trees: the bar chart (Figure 10) shows sorted meteorological variables' importance scores. Next, the dataset containing meteorological data and the corresponding FAO Penman–Monteith ET_0 estimated values are split and fed into the XGBoost model, and the model is then evaluated.

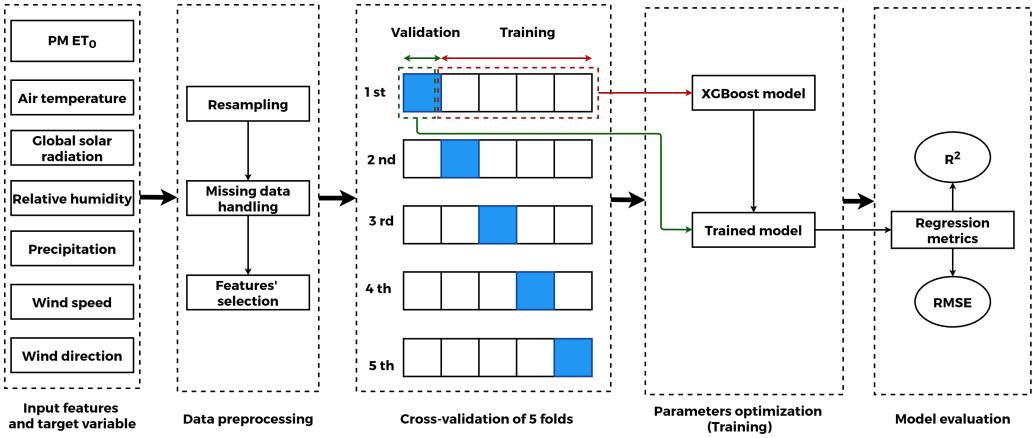

Figure 9. The flowchart of the proposed method.

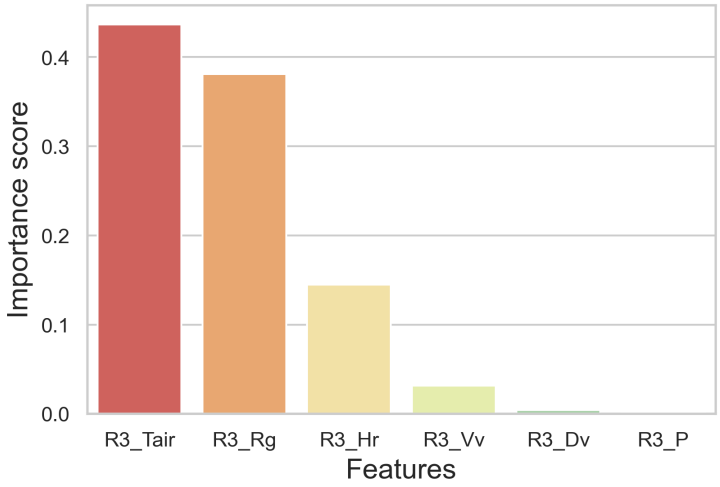

Figure 10. The features' importance bar chart of meteorological parameters.

The objective of this service is to provide an alternative to the FAO Penman–Monteith calculation procedure by learning the behavior of this procedure using a limited number of climatic variables (Figure 11). It is either suitable for stations that lack the necessary hardware and sensors to provide the entire set of meteorological data required for FAO Penman–Monteith or in the case of technical problems with sensors, among other things.

4.4.4. Weather Data Analysis and Visualization Service

It is well-known that data in their raw form are useless, but the information, knowledge, and wisdom derived from them are not. Moving from one state of data to another is known as Knowledge Discovery in Databases (KDD), which is a subset of the modern data science field and can be performed using various methods, such as CRISP-DM [44], which stands for CRoss-Industry Standard Process for Data Mining, or the proposed standard method in [45]. One example of insights data visualization and analysis which can be given in our use case is shown in Figure 12. This is achieved by following a hybrid methodology that includes the following steps: collecting, storing, cleaning, visualizing, analyzing, and mining.

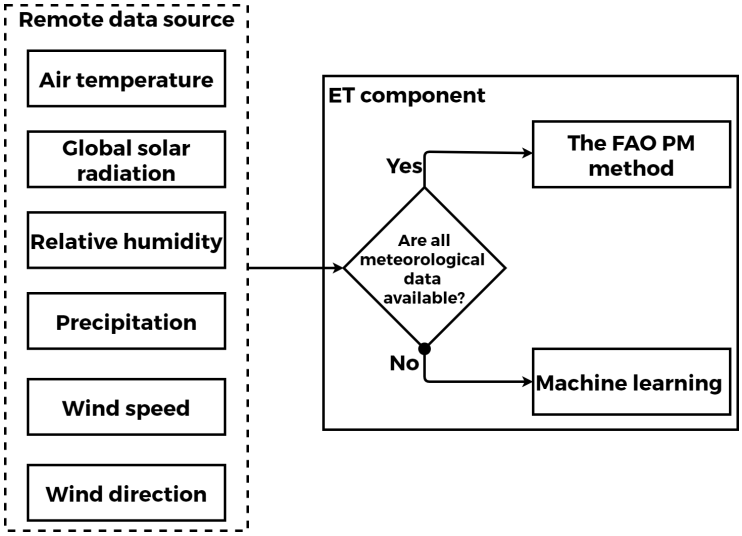

Figure 11. The logic of the Evapotranspiration estimation component.

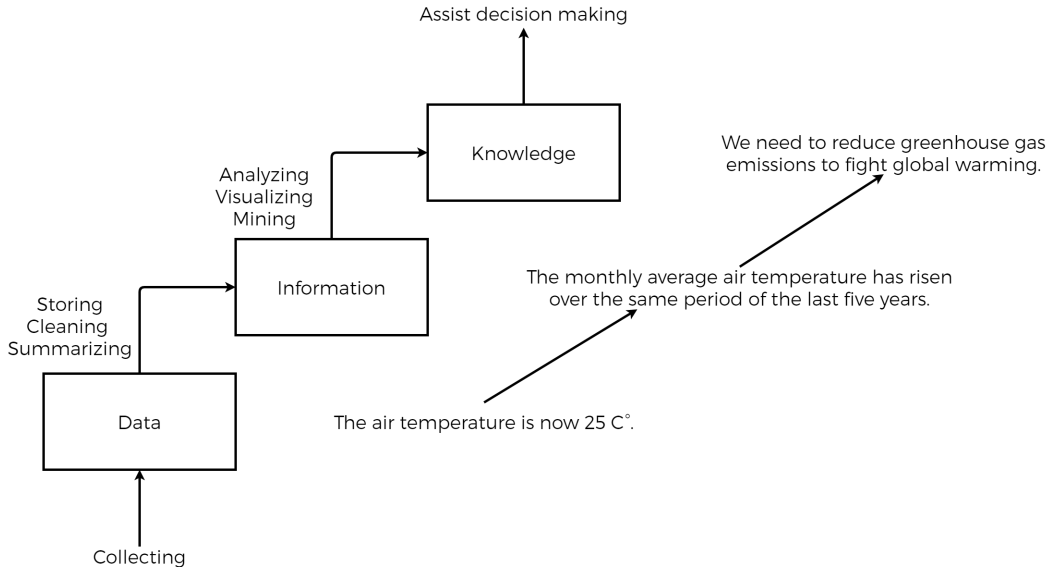

Figure 12. An example of the data analysis scenario.

The first three steps are common for all other services available on the platform. The added value of this service is providing different types of visualization options, including comparison plots (line charts of weather time series), relationship plots (scatter plot of weather data), and automatically generating correlation heat maps, which are important steps in the data analysis phase to study how one variable affects another.

4.4.5. Custom Early Warning Alerts Service

This service is classified as an outlier or anomaly detection problem. These special types of data can be detected in time series by using, among others, rule-based methods, in which case the service alerts administrators via SMS and email once the given condition is satisfied. An example of such a condition is a threshold of temperature or rainfall. The second method of sending warnings is to identify sequences that are notably different from the rest of the historical time series data. These sequences can be the result of measurement error or noise and can inform administrators about the status of the different sensors installed in the meteorological station and inform them about events that could require urgent action. To perform this, we use unsupervised machine learning methods that do not require an annotated series of anomalies to be trained, in contrast to supervised machine learning algorithms. The unsupervised method we used is Local Outlier Factor

(*LOF*) [46]. It uses the KNN algorithm principle to calculate the distance of a point with respect to its neighbors. This distance is used, in turn, to obtain the Reachability Density (*RD*) (Equation (17)).

$$RD(x_i, x_j) = max(K - distance(x_j), distance(x_i, x_j)) \quad (17)$$

Next is measuring the local deviation of the density of a given point using Local Reachability Density (*LRD*) (Equation (18)), which tells us how far the point is from the nearest dense cluster of points.

$$LRD_k(x) = \frac{1}{\sum_{x_j \in N_k(x)} \frac{RD(x,x_j)}{||N_k(x)||}} \quad (18)$$

where $Nk(x)$ is the number of neighbors of x whose distance from x is not greater than the k-distance. As a final step, the algorithm calculates the *LOF* (Equation (19)). Conventionally the points that have a higher anomaly score than their neighbors $LOF > 1$) will be considered as potential outliers, but that is not always true, since in anomaly detection there is no clear and standard validation approach, and the final decision must relies on domain expertise to consider the detected point as an outlier or not, and the role of the service ends by notifying the administrators and letting them decide.

$$LOF_k(x) = \frac{\sum_{x_j \in N_k} LRD(x_j)}{||N_k(x)||} \times \frac{1}{LRD_k(x)} \quad (19)$$

5. Results and Discussions

5.1. Time Series Data Imputation

As a result, for handling missing data using deep learning, Figure 13 shows the curves for the loss function (MSE) and R^2 in both training and validation sets during each epoch. For both architectures (FFNN and LSTM), the learning curves indicate a good fit of the model represented by an initially high training loss that steadily decreases as more training instances are added and flattens over time (0.04 for air temperature, 0.098 for global solar radiation, and 0.116 for air relative humidity), and the same way for R^2, which begins with nonoptimal values in training and validation and converges to stable change (0.96 for air temperature, 0.84 for global solar radiation, and 0.77 for air relative humidity). Table 5 shows the performance comparisons between the two deep learning architectures used in the test set.

Table 5 shows the performance comparisons between the two deep learning models used in the test set.

Table 5. Performance of deep learning models.

Metric/Model	FFNN			LSTM		
	R3_Tair	***R3_Rg***	***R3_Hr***	***R3_Tair***	***R3_Rg***	***R3_Hr***
Training time (s)	68.371	34.943	60.639	138.193	65.574	178.503
R^2	0.957	0.838	0.768	0.957	0.839	0.812
R	0.978	0.916	0.877	0.978	0.916	0.901
RMSE	0.037	0.098	0.116	0.037	0.097	0.105
MAE	0.029	0.069	0.094	0.029	0.066	0.081

Once trained, the final deep learning model will be ready for deployment in the production environment and used to make inferences about meteorological station data given the ERA5-Land data.

In contrast to Numerical Methods for Weather Prediction (NWP), which take considerable time to run [47], predictions based on our model are made instantly.

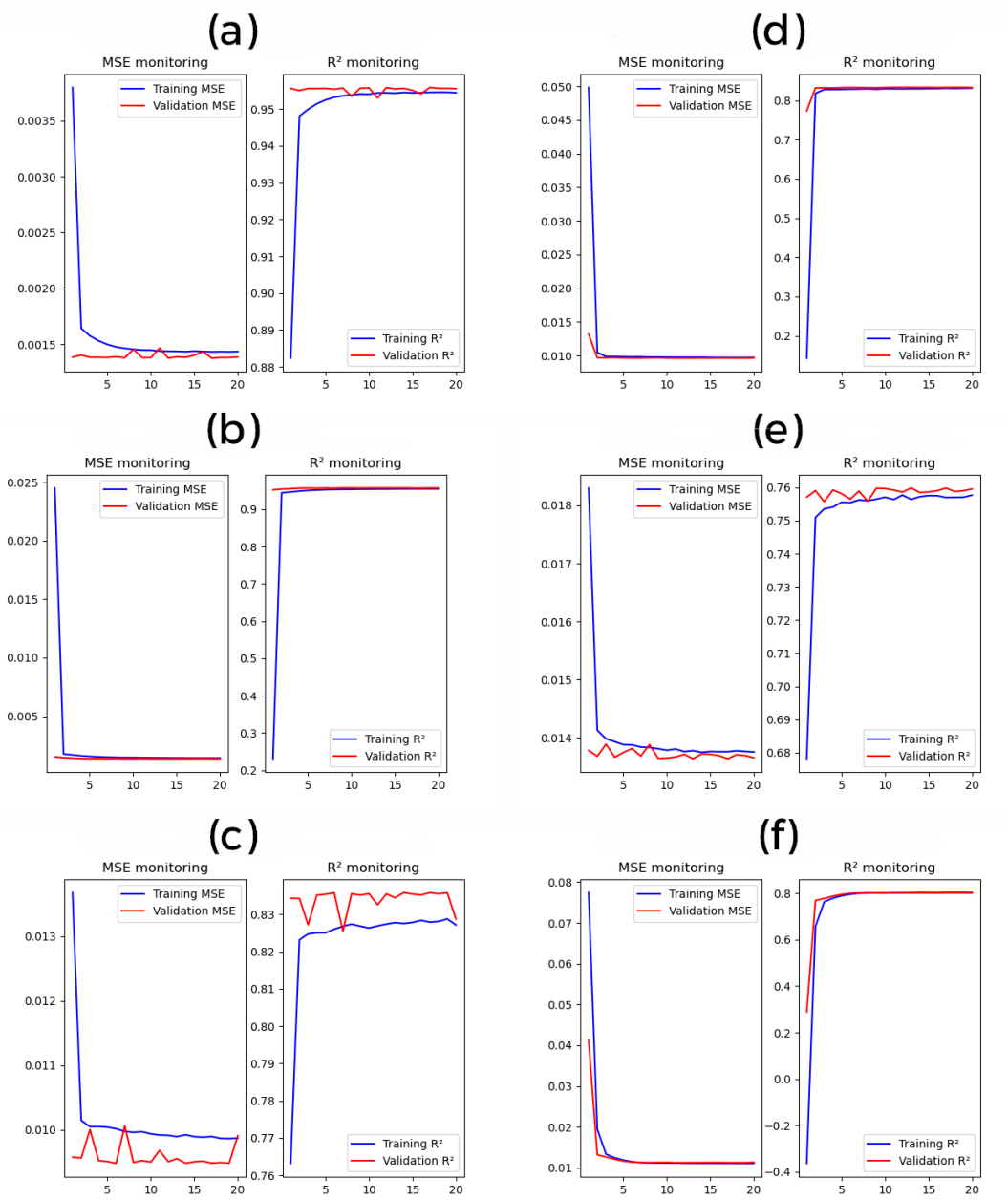

Figure 13. Monitoring of MSE and R^2 during training and validation phases: air temperature: (**a**) FFNN, (**b**) LSTM, global solar radiation: (**c**) FFNN, (**d**) LSTM, air relative humidity: (**e**) FFNN, (**f**) LSTM.

5.2. Climatic Parameters Calculation and Estimation

We split our dataset into five randomly shuffled parts (five folds). We used one fold for model evaluation and the remaining four to train it. Finally, we assess the model's performance by calculating the regression metrics across the five folds of the dataset (Root Mean Squared Error RMSE and the coefficient of determination R^2). According to the results of Table 6, the main point is not the perfect results obtained by using all parameters as estimators but using only air temperature and global solar radiation (average RMSE = 0.27 and average R^2 = 0.93), which gave promising results. This represents a practical data-driven approach for accurately estimating ET_0 by combining two criteria: accuracy of the FAO Penman–Monteith method and using limited and simple-to-obtain meteorological variables (air temperature with global solar radiation or only air temperature).

Table 6. Cross-validation results for the three scenarios.

Fold	All Variables		*R3_Tair* and *R3_Rg*		Only *R3_Tair*	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	0.976094	0.080174	0.922928	0.258481	0.727668	0.913333
2	0.978442	0.092729	0.922629	0.332801	0.759759	1.033369
3	0.981453	0.066742	0.942489	0.206956	0.760640	0.861347
4	0.978672	0.085282	0.938638	0.245365	0.770725	0.916799
5	0.979803	0.081603	0.930292	0.281652	0.757509	0.979770

According to the results, the proposed method represents a practical data-driven approach for accurately estimating ET_0 by combining two criteria: accuracy of the FAO Penman–Monteith method and using limited and simple-to-obtain meteorological variables (air temperature with global solar radiation or only air temperature).

The results obtained are in agreement with the approach followed in [48] for the most contributing variables for ET_0 estimation, but it outperforms it in terms of RMSE (0.19). In addition, the XGboost model used in this study outperforms [49] when using all meteorological variables to train a neural network (RMSE = 0.19).

5.3. Prototype of the System

The platform is named "FLA7A", which means agriculture in Arabic. Figure 14 shows a screenshot of the platform's dashboard. By default, it contains the real-time visualization of the last three days of hourly weather data (line charts). However, the user can customize the period according to their needs.

Figure 14 shows a screenshot of the dashboard that visualizes the last three days of hourly weather data.

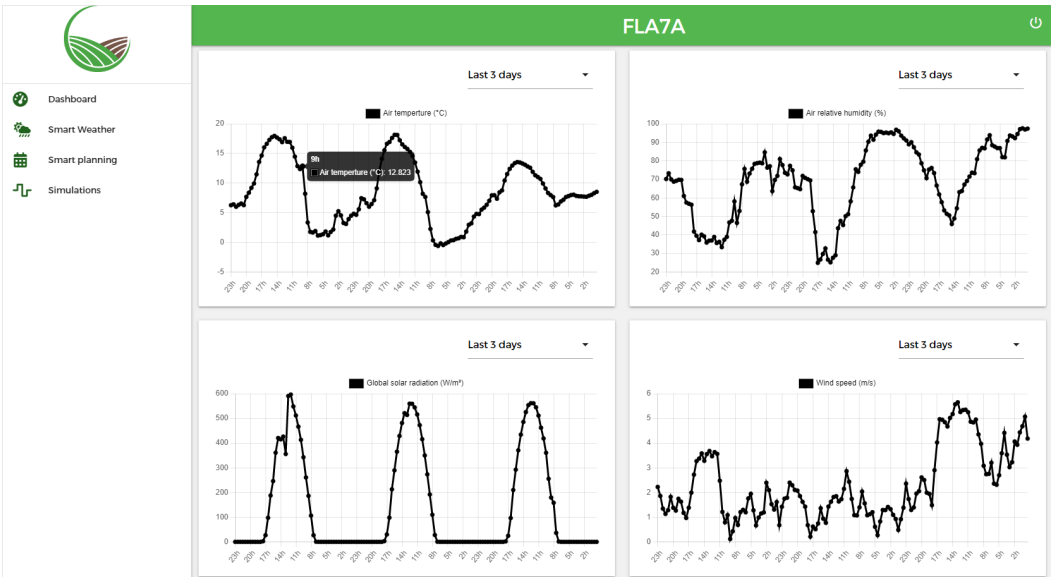

Figure 14. A screenshot of the platform's real-time weather time series visualization service. The black dots represent the original measurements, while the black line represents the linear interpolation of the dots.

Figure 15 shows the interface of the forecasting service. By default, the forecasting period is set to one year. However, the user can change this parameter and also customize the number of years it will take into consideration when training the model. As mentioned in Section 4.4.2, the service initially uses the statistical model Facebook Prophet, which is powerful in long-term forecasting. In future work, the platform will integrate more models,

especially those known for their performance in the task of short-term and mid-term weather forecasting, as surveyed in [50].

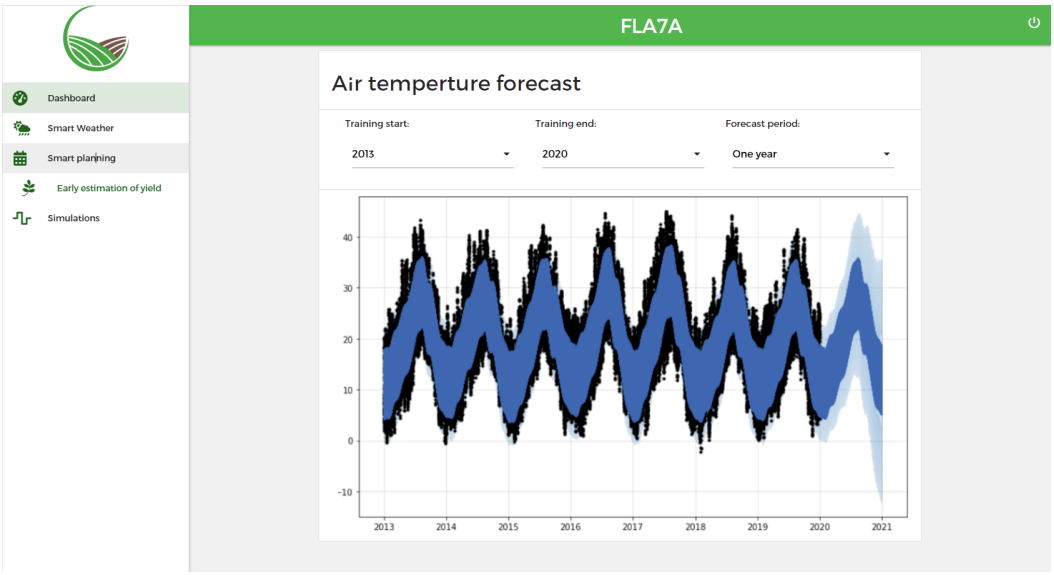

Figure 15. A screenshot of the platform's forecast service. The light blue is the uncertainty bounds of the uncertainty interval around the final predictions (upper and lower), while the dark blue is the predicted values, and the black dots represent our original data.

This platform is the first incremental step towards our goal of creating a decision support system intended to implement smart agriculture in Morocco by using artificial intelligence and data science to solve real-life problems facing farmers. More studies will be performed to add and optimize different parts of this system. For example, for data storage, we used MongoDB as a database to deal with real-time big data, however, in terms of system scalability or dealing with batch processing or long-running ETL (Extract, Transform, and Load) jobs, other technologies should be considered. A potential candidate for this could be other NoSQL databases [51] or the Hadoop ecosystem (Spark, MapReduce, Hive, etc.) [52,53]. Additionally, MongoDB has fault tolerance issues, which is true of practically all distributed databases. Moreover, in the proposed deep learning method to deal with missing data, we do not cover hyperparameter fine-tuning [54], which is one of the biggest drawbacks to using deep neural networks. This task can be performed using GridSearch [55] or other optimization algorithms such as Genetic Algorithms or the Monte Carlo method, which can lead to better results. Despite this, we found promising results, confirming the reliability of the ERA5-Land reanalysis data for our study area, which could lead to the application of this method using several stations in regions with challenging conditions. Finally, for the anomaly detection part, we presented an unsupervised machine learning method that is based on how isolated a measure is with respect to the surrounding neighborhood to alert administrators via SMS and emails, but other efficient methods can be investigated, such as Conformal Anomaly Detection (CAP) [56,57].

6. Conclusions

The work conducted in this paper makes use of artificial intelligence and big data analytics to develop a platform intended for intelligent weather data management, which is essential to implementing smart and sustainable agriculture in Morocco. The platform exploits huge amounts of generated data to gain insights and help assist decision making. The proposed platform offers a number of weather-data-related services such as handling missing data, visualization, analysis, estimation, and forecasting. Combining ERA5-Land reanalysis data and deep learning algorithms to learn the relationship between the two data sources gave promising results (R^2 = 0.96 and RMSE = 0.04) for the air temperature

variable. The same is true for the estimation of the reference evapotranspiration using XGBoost (R^2 = 0.97 and RMSE = 0.07). The platform is designed to be service-oriented and will incorporate other services and solutions to help farmers and policymakers.

Author Contributions: Platform and machine learning models development, C.E.H. and S.B.; data acquisition and processing, C.E.H., S.B., B.S. and S.K.; methodology, C.E.H., S.B., S.K. and A.C.; writing—original draft, C.E.H.; writing—review and editing, S.B., S.K., D.D. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: This study was supported by and conducted within the Center for Remote Sensing Applications (CRSA) at the Mohammed VI Polytechnic University (UM6P) in Morocco. (https://crsa.um6p.ma/ (accessed on 1 September 2022)).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wade, M.; Hoelle, J.; Patnaik, R. Impact of Industrialization on Environment and Sustainable Solutions—Reflections from a South Indian Region. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *120*, 012016. [CrossRef]
2. Bongaarts, J. Human population growth and the demographic transition. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 2985. [CrossRef] [PubMed]
3. Doungmanee, P. The nexus of agricultural water use and economic development level. *Kasetsart J. Soc. Sci.* **2016**, *37*, 38–45. [CrossRef]
4. Frisvold, G.; Sanchez, C.; Gollehon, N.; Megdal, S.B.; Brown, P. Evaluating Gravity-Flow Irrigation with Lessons from Yuma, Arizona, USA. *Sustainability* **2018**, *10*, 1548. [CrossRef]
5. Belaqziz, S.; Mangiarotti, S.; Le Page, M.; Khabba, S.; Er-Raki, S.; Agouti, T.; Drapeau, L.; Kharrou, M.H.; El Adnani, M.; Jarlan, L. Irrigation scheduling of a classical gravity network based on the Covariance Matrix Adaptation—Evolutionary Strategy algorithm. *Comput. Electron. Agric.* **2014**, *102*, 64–72. [CrossRef]
6. Nafchi, R.A. Evaluation of the Efficiency of the Micro-irrigation Systems in Gardens of Chaharmahal and Bakhtiari Province of Iran. *Int. J. Agric. Econ.* **2021**, *6*, 106–110. [CrossRef]
7. Norasma, C.Y.N.; Fadzilah, M.A.; Roslin, N.A.; Zanariah, Z.W.N.; Tarmidi, Z.; Candra, F.S. Unmanned Aerial Vehicle Applications In Agriculture. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *506*, 012063. [CrossRef]
8. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
9. Rienecker, M.M.; Suarez, M.J.; Gelaro, R.; Todling, R.; Bacmeister, J.; Liu, E.; Bosilovich, M.G.; Schubert, S.D.; Takacs, L.; Kim, G.K.; et al. MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Clim.* **2011**, *24*, 3624–3648. [CrossRef]
10. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H.; et al. The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *J. Meteorol. Soc. Jpn. Ser. II* **2015**, *93*, 5–48. [CrossRef]
11. Kanamitsu, M.; Ebisuzaki, W.; Woollen, J.; Yang, S.K.; Hnilo, J.J.; Fiorino, M.; Potter, G.L. NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **2002**, *83*, 1631–1644. [CrossRef]
12. Majumdar, P.; Mitra, S. IoT and Machine Learning-Based Approaches for Real Time Environment Parameters Monitoring in Agriculture: An Empirical Review. *Agric. Inform.* **2021**, *5*, 89–115. [CrossRef]
13. Kumar, S.; Ansari, M.A.; Pandey, S.; Tripathi, P.; Singh, M. Weather Monitoring System Using Smart Sensors Based on IoT. *Lect. Notes Netw. Syst.* **2020**, *106*, 351–363._36. [CrossRef]
14. Kodali, R.K.; Mandal, S. IoT Based Weather Station. In Proceedings of the 2016 International Conference on Control Instrumentation Communication and Computational Technologies, ICCICCT 2016, Kumaracoil, India, 16–17 December 2016; pp. 680–683. [CrossRef]
15. Mittal, Y.; Mittal, A.; Bhateja, D.; Parmaar, K.; Mittal, V.K. Correlation among Environmental Parameters Using an Online Smart Weather Station System. In Proceedings of the 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015, Delhi, India, 17–20 December 2015. [CrossRef]
16. Djordjević, M.; Jovičić, B.; Marković, S.; Paunović, V.; Danković, D. A smart data logger system based on sensor and Internet of Things technology as part of the smart faculty. *J. Ambient Intell. Smart Environ.* **2020**, *12*, 359–373. [CrossRef]
17. Amin, F.; Abbasi, R.; Mateen, A.; Ali Abid, M.; Khan, S. A Step toward Next-Generation Advancements in the Internet of Things Technologies. *Sensors* **2022**, *22*, 8072. [CrossRef] [PubMed]

18. Kamilaris, A.; Kartakoullis, A.; Prenafeta-Boldú, F.X. A review on the practice of big data analysis in agriculture. *Comput. Electron. Agric.* **2017**, *143*, 23–37. [CrossRef]
19. Muangprathub, J.; Boonnam, N.; Kajornkasirat, S.; Lekbangpong, N.; Wanichsombat, A.; Nillaor, P. IoT and agriculture data analysis for smart farm. *Comput. Electron. Agric.* **2019**, *156*, 467–474. [CrossRef]
20. Math, R.K.M.; Dharwadkar, N.V. IoT Based low-cost weather station and monitoring system for precision agriculture in India. In Proceedings of the 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Palladam, India, 30–31 August 2018; pp. 81–86. [CrossRef]
21. Djordjevic, M.; Dankovic, D. A Smart Weather Station Based on Sensor Technology. *Facta Univ. Ser. Electron. Energetics* **2019**, *32*, 195–210. [CrossRef]
22. Roukh, A.; Fote, F.N.; Mahmoudi, S.A.; Mahmoudi, S. WALLeSMART: Cloud Platform for Smart Farming. In Proceedings of the 32nd International Conference on Scientific and Statistical Database Management, Vienna, Austria, 7–9 July 2020. [CrossRef]
23. Er-Raki, S.; Chehbouni, A.; Duchemin, B. Combining Satellite Remote Sensing Data with the FAO-56 Dual Approach for Water Use Mapping In Irrigated Wheat Fields of a Semi-Arid Region. *Remote Sens.* **2010**, *2*, 375–387. [CrossRef]
24. Belaqziz, S.; Khabba, S.; Kharrou, M.H.; Bouras, E.H.; Er-Raki, S.; Chehbouni, A. Optimizing the Sowing Date to Improve Water Management and Wheat Yield in a Large Irrigation Scheme, through a Remote Sensing and an Evolution Strategy-Based Approach. *Remote Sens.* **2021**, *13*, 3789. [CrossRef]
25. Er-Raki, S.; Chehbouni, A.; Guemouria, N.; Duchemin, B.; Ezzahar, J.; Hadria, R. Combining FAO-56 model and ground-based remote sensing to estimate water consumptions of wheat crops in a semi-arid region. *Agric. Water Manag.* **2007**, *87*, 41–54. [CrossRef]
26. El Hachimi, C.E.; Belaqziz, S.; Khabba, S.; Chehbouni, A. Towards Precision Agriculture in Morocco: A Machine Learning Approach for Recommending Crops and Forecasting Weather. In Proceedings of the 2021 International Conference on Digital Age and Technological Advances for Sustainable Development, ICDATA 2021, Marrakech, Morocco, 29–30 June 2021; pp. 88–95. [CrossRef]
27. Aouade, G.; Ezzahar, J.; Amenzou, N.; Er-Raki, S.; Benkaddour, A.; Khabba, S.; Jarlan, L. Combining stable isotopes, Eddy Covariance system and meteorological measurements for partitioning evapotranspiration, of winter wheat, into soil evaporation and plant transpiration in a semi-arid region. *Agric. Water Manag.* **2016**, *177*, 181–192. [CrossRef]
28. Kharrou, M.H.; Simonneaux, V.; Er-Raki, S.; Le Page, M.; Khabba, S.; Chehbouni, A. Assessing Irrigation Water Use with Remote Sensing-Based Soil Water Balance at an Irrigation Scheme Level in a Semi-Arid Region of Morocco. *Remote Sens.* **2021**, *13*, 1133. [CrossRef]
29. Oses, N.; Azpiroz, I.; Marchi, S.; Guidotti, D.; Quartulli, M.; Olaizola, I.G. Analysis of Copernicus' ERA5 Climate Reanalysis Data as a Replacement for Weather Station Temperature Measurements in Machine Learning Models for Olive Phenology Phase Prediction. *Sensors* **2020**, *20*, 6381. [CrossRef] [PubMed]
30. Zandler, H.; Senftl, T.; Vanselow, K.A. Reanalysis datasets outperform other gridded climate products in vegetation change analysis in peripheral conservation areas of Central Asia. *Sci. Rep.* **2020**, *10*, 22446. [CrossRef]
31. Bui, M.T.; Lu, J.; Nie, L. Evaluation of the Climate Forecast System Reanalysis data for hydrological model in the Arctic watershed Målselv. *J. Water Clim. Chang.* **2021**, *12*, 3481–3504. [CrossRef]
32. Muñoz-Sabater, J.; Dutra, E.; Agustí-Panareda, A.; Albergel, C.; Arduini, G.; Balsamo, G.; Boussetta, S.; Choulga, M.; Harrigan, S.; Hersbach, H.; et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **2021**, *13*, 4349–4383. [CrossRef]
33. ERA5-Land Hourly Data from 1950 to Present. Available online:. (accessed on 1 September 2022) [CrossRef]
34. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597. [CrossRef]
35. El Hachimi, C.; Belaqziz, S.; Khabba, S.; Chehbouni, A. Data Science Toolkit: An all-in-one python library to help researchers and practitioners in implementing data science-related algorithms with less effort. *Softw. Impacts* **2022**, *1*, 100240. [CrossRef]
36. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
37. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
38. El Hachimi, C.; Belaqziz, S.; Khabba, S.; Chehbouni, A. Early Estimation of Daily Reference Evapotranspiration Using Machine Learning Techniques for Efficient Management of Irrigation Water. *J. Phys. Conf. Ser.* **2022**, *2224*, 012006. [CrossRef]
39. Lawrence, M.G. The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications. *Bull. Am. Meteorol. Soc.* **2005**, *86*, 225–234. [CrossRef]
40. Parish, O.; Putnam, T.W. *NASA Equations for the Determination of Humidity from Dewpoint and Psychrometric Data*; NASA Dryden Flight Research Center: Hampton, VA, USA, 1977.
41. Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **1998**, *6*, 107–116. [CrossRef]
42. Bhanja, S.; Das, A. Impact of Data Normalization on Deep Neural Network for Time Series Forecasting. *arXiv* **2018**. [CrossRef]
43. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [CrossRef]

44. Wirth, R.; Wirth, R. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery And Data Mining, Denham, UK, 11–13 April 2000; pp. 29–39.
45. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
46. Breuniq, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD Record* **2000**, *29*, 93–104. [CrossRef]
47. Carroll, A.B.; Wetherald, R.T. Application of Parallel Processing to Numerical Weather Prediction. *J. ACM* **1967**, *14*, 591–614. [CrossRef]
48. Pal, M.; Deswal, S. M5 model tree based modelling of reference evapotranspiration. *Hydrol. Process.* **2009**, *23*, 1437–1443. [CrossRef]
49. Yassin, M.A.; Alazba, A.A.; Mattar, M.A. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. *Agric. Water Manag.* **2016**, 163, 110–124. [CrossRef]
50. Schultz, M.G.; Betancourt, C.; Gong, B.; Kleinert, F.; Langguth, M.; Leufen, L.H.; Mozaffari, A.; Stadtler, S. Can Deep Learning Beat Numerical Weather Prediction? *Philos. Trans. R. Soc. A* **2021**, *379*, 20200097. [CrossRef] [PubMed]
51. Gessert, F.; Wingerath, W.; Friedrich, S.; Ritter, N. NoSQL database systems: A survey and decision guidance. *Comput. Sci.—Res. Dev.* **2016**, *32*, 353–365. [CrossRef]
52. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop Distributed File System. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010, Incline Village, NV, USA, 3–7 May 2010. [CrossRef]
53. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache spark: A unified engine for big data processing. *Commun. ACM* **2016**, *59*, 56–65. [CrossRef]
54. Probst, P.; Bischl, B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *J. Mach. Learn. Res.* **2019**, *20*, 1–32.
55. Bergstra, J.; Ca, J.B.; Ca, Y.B. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
56. Nouretdinov, I. Distributed Conformal Anomaly Detection. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 253–258. [CrossRef]
57. Laxhammar, R.; Falkman, G. Online detection of anomalous sub-trajectories: A sliding window approach based on conformal anomaly detection and local outlier factor. *IFIP Adv. Inf. Commun. Technol.* **2012**, *382*, 192–202. [CrossRef]