Hybrid Machine Learning Approach for Evapotranspiration Estimation of Fruit Tree in Agricultural Cyber-Physical Systems

Tianteng Wang, Xuping Wang, Yiping Jiang, Zilai Sun, Yuhu Liang, Xiangpei Hu, Hao Li, Yan Shi, Jingjun (David) Xu, and Junhu Ruan

Abstract—The flourish of the Internet of Things (IoT) and data-driven techniques provide new ideas for enhancing agricultural production, where evapotranspiration estimation is a crucial issue in crop irrigation systems. However, tremendous and unsynchronized data from agricultural cyber-physical systems bring large computational costs as well as complicating performing conventional machine learning methods. To precisely estimate evapotranspiration with acceptable computational costs under the background of IoT, we combine time granulation computing techniques and Gradient Boosting Decision Tree (GBDT) with Bayesian Optimization (BO) to propose a hybrid machine learning approach. In the combination, a fuzzy granulation method and a time calibration technique are introduced to reduce voluminous and unsynchronized data into small-scale and synchronized granules with high representativeness. Subsequently, GBDT is implemented to predict evapotranspiration, and BO is utilized to find the optimal hyper-parameter values from the reduced granules. IoT data from Xi’an Fruit Technology Promotion Center in Shaanxi Province, China verify that the proposed granular-GBDT-BO is effective for cherry tree evapotranspiration estimation with reduced computational time, acceptable and robust predictive accuracy. Consequently, the precise estimation of crop evapotranspiration could provide operational guidance for plant irrigation, plant conservations, and pest control in the agricultural greenhouse.

Index Terms—Machine learning, agricultural cyber-physical systems, fuzzy system, and agricultural greenhouse

I. INTRODUCTION

The explosion of global population, the diminishing arable lands and the increasing demands for foods have appealed for industrialization and intensification of agriculture [1]. The “Internet of Things (IoT)” is a highly promising family of techniques used to achieve the digitalization and modernization of agriculture [2], [3]. In 2021, the global agricultural internet of things (AgIoT) market value starts at approximately 4.02 billion U.S. dollars and is anticipated to grow over the next few years, reaching almost 7 billion U.S. dollars by 2025 [36]. In academia, some special issues from journals also call for papers for further investigations on agricultural IoT, as shown in [38], [39]. Specifically, IoT applications can enhance agricultural production and operations, encompassing yield prediction, disease detection, seed detection, livestock management (i.e., animal welfare and livestock production), and soil management [4]. For example, drought rural areas forced farmers to use less water. And smart sprinkler systems (i.e., actuators) which monitor real-time conditions could automatically trigger the irrigation on when the soil humidity goes below the desired thresholds without human interventions, which enables farmers to cut water and fertilizer utilization by up to 40 percent without reducing yields, according to Nature Conservancy [37].

Water management in agricultural production plays a paramount role in hydrological, climatological, and agronomical balance [6]. In water management, the accurate evapotranspiration estimation is vital for irrigation systems [4]. Evapotranspiration is defined as the transportation of water from crop root to leaf surface, which identifies the total amount of water lost from a crop. This process is highly complicated, which includes the interaction of various plants, atmosphere, and soil elements. There exist three streams for evapotranspiration estimation. First and traditionally, evapotranspiration is measured by using experimental techniques (such as lysimeters, Bowen ratio energy balance, and eddy covariance systems) [7], [8], [9], but higher operating costs and lower accuracy of these physical devices have restricted these methods in practice. Second and subsequently, numerous equations are proposed to be another alternative. For example, the Food and Agricultural Organization of United Nations treated the Penman-Monteith equation (FAO-56PM) as a standard estimation method in 1998 [10]. Jensen et al., proposed temperature-based Hargreaves (HG) equation for estimation when data are not fully available [11]. However, mathematical equations have strong assumptions that real-life agriculture always cannot comply with. Third, machine learning models have provided new solutions in the recent years. Mehdiabadeh et al., implemented Support Vector Machine (SVM) and multivariable adaptive regression splines (MARS) to estimate monthly evapotranspiration in arid and semi-arid regions with 264 data records from 22 meteorological stations [12]. Feng et al., estimated the daily reference evapotranspiration in six meteorological stations by using a
generalized regression neural network in Sichuan, China with data from 1961 to 2014 [13]. Patil et al., used extreme learning to estimate weekly evapotranspiration in the Thar Desert, India with 1920 data records [14]. Recently, the development of cyber-physical systems in the agricultural sector provides tremendous and extensive data resources, which grant new research opportunities for data-driven research. As such, the fluctuation of cyber-physical system data could accurately capture crop in real time. Many scholars have made meaningful explorations regarding cyber-physical systems data (i.e., streaming data). Yu et al., propose a novel continuous support vector regression for nonstationary streaming data [15]. Kowsar et al., come up with an online unsupervised dynamic window method to track repeating patterns from sensor data [16]. However, different from past machine learning-based research, two key problems exist when performing machine learning models for evapotranspiration estimation under the background of agricultural cyber-physical systems. One is that continuous and tremendous data are collected from diverse IoT equipment. High frequent and massive IoT data may limit the applications of conventional machine learning methods, as reviewed in [4]. The other is that heterogeneous sensors generate different data records at one fixed period (i.e., different data updating frequencies), hence the features (several inputs) and target (one output) of machine learning models are not one-by-one synchronized. In this situation, it is almost impossible to perform machine learning models and make estimations. Motivated by these observations, one of our contributions is to combine granulation computing with time calibration technique to reduce voluminous and unsynchronized data into small-scale and synchronized data with high representativeness.

From the perspective of methodology, the estimation (i.e., prediction) of plant evapotranspiration is a regression problem in essence. As for regression problems in agriculture-oriented and sensor-based problems, SVM and neural network (NN) are two principal and widely used supervised learning algorithms in past literature [17], [18]. Yu et al., design an online robust support vector regression for data streams [19]. Vasconcelos et al., propose a mixed Gaussian regression model for bimodal agriculture data [20]. Yu et al., bring forward a topology learning-based fuzzy random neural network for streaming data regression [21]. However, SVM cannot handle voluminous cyber-physical systems data well due to long training time and NN is weak in explanatory/interpretation, as well as easily traps in local minima, which limit their further applications in agricultural cyber-physical systems. Gradient Boosting Decision Tree (GBDT) is one recently developed ensemble learning method which has been widely used in academia and industry due to its state-of-the-art performance in many machine learning tasks [22]. Ma et al., [23] proposed a GBDT-based method for transportation accident causal analysis and prediction. Wen et al., [24] presented a novel parallel implementation for GBDT on Graphics Processing Units (GPUs). Fröhlich et al., [25] utilized a GBDT-based method to establish a computational gait analysis approach for the Noldus Catwalk system. Liu et al., [26] presented a geomagnetic data reconstruction approach based on Gradient Boosting Decision Tree. However, two shortcomings are obvious in these studies. One is that the performance of conventional GBDT is closely related to the initial setting of hyper-parameter, but no definite analytical expression exists in GBDT prediction function (i.e., black-box) [22]. Therefore, conventional Gradient and Hessian matrix-based optimization methods are inferior for this type of problem. The other is that conventional GBDT faces immense computational costs when it meets voluminous agricultural cyber-physical systems data [22]. To surmount these problems, we integrate granulation technique, time calibration technique, and GBDT with BO to propose a granular-GBDT-BO approach for crop’s evapotranspiration estimation in greenhouse. The granulation technique can effectively break voluminous data into small-scale representative data and BO helps GBDT in obtaining optimal hyper-parameter values.

Ultimately, this paper makes the following contributions. First, we formulate a framework for crop evapotranspiration estimation in agricultural greenhouse cyber-physical systems from data-driven perspective, which complements evapotranspiration estimation under the background of agricultural greenhouse cyber-physical systems. ii) A granular-GBDT-BO approach for agricultural cyber-physical systems data is proposed, where fuzzy granulation and time calibration techniques are presented to break voluminous and unsynchronized data into small-scale synchronized granules with equivalent representativeness of original data. Consequently, the training time of granular-GBDT-BO has greatly reduced with satisfactory predictive accuracy.

The remaining of our work is organized as follows. Section II introduces a framework of evapotranspiration estimation in agricultural greenhouse cyber-physical systems and IoT dataset. Section III proposes a granular-GBDT-BO approach, comprising a time calibration granular method and a GBDT-BO algorithm. Section IV shows the results and corresponding analysis, verifying the superiorities of the work. Section V concludes the work.

II. FRAMEWORK OF AGRICULTURAL CYBER-PHYSICAL SYSTEMS AND USED IOT DATA

A. Framework of Estimating Evapotranspiration in Agricultural Cyber-Physical Systems

On the basis of the experience of other scholars [27], [28], we formulate a complementary three-layer framework for crop evapotranspiration estimation in agricultural cyber-physical systems, as shown in Fig. 1. They are introduced as follows:

a) Perception layer. This layer encompasses various sensors utilized in agriculture to collect agricultural data and perceive the state of crops and greenhouse environment. This layer includes but is not limited to temperature and humidity sensors, light sensors, physical devices, RFID, GPS, meteorological equipment, etc. As for the problem of crop’s evapotranspiration estimation, air temperature, air humidity, soil temperature, soil moisture, the concentration of carbon dioxide and effective illumination radiation data are collected.

b) Computing layer. The computing layer includes storage and computing devices. Original agriculture data are pre-processed by time calibration granular technique. Moreover, a machine learning-based algorithm (i.e., GBDT) is
applied for crop’s evapotranspiration estimation, and GBDT is improved by BO.

c) Application layer. Five main agricultural application scenarios are illustrated in this layer, including yield prediction, disease detection, species recognition, livestock production, and evapotranspiration estimation. With the deep integration of interdisciplinary interactions among computer science, agriculture science, production expenses will diminish with facilitated production efficiency.

Crop’s evapotranspiration at a period of time is defined as stem runoff. Therefore, precise estimation of stem runoff is the paramount step for crop’s evapotranspiration estimation.

B. Dataset of Agricultural Cyber-Physical Systems

This study uses the IoT data from the information platform at Xi’an Fruit Technology Promotion Center, which is located in Xi’an City, Shaanxi Province, China. The dataset is available at http://113.200.93.215/Portal/Xagy.aspx. Cherry is a special and widely planted fruit in Shaanxi Province, China, and the data source of the cherry tree is complete, thus we focus on cherry tree greenhouse. The greenhouse is a relatively closed ecosystem, thus, its inner environment has significantly and positively impact the stem runoff of cherry trees, as shown in Fig. 2. Therefore, we use greenhouse environmental data to predict the stem runoff of cherry trees.

Environmental data in greenhouse represent the ecological information of greenhouse, containing air temperature, air humidity, soil temperature, soil moisture, the concentration of CO₂ and effective illumination radiation. These data update simultaneously in an irregular frequency due to the sensors. They sometimes update once per 4 minutes and sometimes update once per 6 minutes. Greenhouse environmental data are collected from December 23, 2018 to December 23, 2019, including 66,634 data records with 6 features, as shown in Figs. 4-9. The basic statistical characteristics of environmental data are elaborated in TABLE I.

Different from greenhouse environmental data, cherry stem runoff updates once per 10 minutes regularly. Simultaneously, stem runoff data are collected from December 23, 2018 to December 23, 2019, including 49,055 records, as shown in Fig. 10. The data distribution is illustrated in Fig. 3, where the unit of stem runoff is gram per square meter per second (g/m²/s).
TABLE I
BASIC STATISTICAL INFORMATION OF ENVIRONMENTAL DATA IN GREENHOUSE.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>STD</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air temperature (°C)</td>
<td>10.52</td>
<td>-9.80</td>
<td>15.62</td>
<td>46.10</td>
<td>-0.13</td>
<td>-0.62</td>
</tr>
<tr>
<td>Air humidity (%)</td>
<td>22.25</td>
<td>10.00</td>
<td>72.40</td>
<td>99.00</td>
<td>-0.58</td>
<td>-0.74</td>
</tr>
<tr>
<td>Soil temperature (°C)</td>
<td>7.25</td>
<td>-15.9</td>
<td>15.50</td>
<td>120.80</td>
<td>-0.34</td>
<td>0.44</td>
</tr>
<tr>
<td>Soil moisture (%)</td>
<td>4.77</td>
<td>0.00</td>
<td>45.13</td>
<td>99.00</td>
<td>-3.34</td>
<td>34.49</td>
</tr>
<tr>
<td>CO₂ density (ppm)</td>
<td>50.41</td>
<td>54.00</td>
<td>423.7</td>
<td>654.00</td>
<td>-0.26</td>
<td>3.16</td>
</tr>
<tr>
<td>Illumination (μmol/m²/s)</td>
<td>71.89</td>
<td>0.00</td>
<td>35.32</td>
<td>838.00</td>
<td>3.80</td>
<td>21.28</td>
</tr>
</tbody>
</table>

Based on descriptive statistical data analysis in Table I and Figs. 4-10, we find these two categories of data have massive but different data records at the same period. To be more specific, environmental data have 6 features with 66,634 data records but the stem runoff data have 1 feature with 49,055 data records. These increasing data bring computing pressures for conventional machine learning algorithms.

Moreover, cherry runoff data and greenhouse environmental data are not one-by-one corresponded (i.e., timestamp is not consistent), as shown in Fig. 2 (i.e., greenhouse environmental data have 10 records, whereas stem runoff data have five
records within one hour at 17-Octotor-2019). The reason is because heterogeneous sensors have different updating frequencies, leading to different data records at one fixed period. However, the machine learning-based research paradigm assumes that features and targets should have the same data records with one-by-one correspondence. Thus, this situation appeals for advanced methods to handle these voluminous and unsynchronized agriculture cyber-physical systems data and make predictions at acceptable costs.

III. METHODOLOGY

In this section, we formulate a time calibration granular technique to break down voluminous and unsynchronized data into small-scale and synchronized granulates, and then integrate the granulation process into GBDT and BO. Fig. 11 depicts a rough overview of our methodology.

A. Time Calibration Granular Technique

The original agricultural cyber-physical systems data in Section II. B is represented by: $X'_{\text{orig}} = \{x'_1, x'_2, \cdots, x'_i\}$, where $i = 1, 2, \cdots, 7$ stands for greenhouse environmental data and cherry stem runoff data. $t$ represents the total records of each variable. First, we divide the original data $X'_{\text{orig}}$ into small-scale sub-data $X'_{\text{subgrp}}$:

$$X'_{\text{subgrp}} = \{x'_1, x'_2, \cdots, x'_i\}, \{x'_{i+1}, x'_{i+2}, \cdots, x'_{i+w}\}, \cdots, \{x'_{i+j+w-3}, x'_{i+j+w-2}, \cdots, x'_{i+j+w}\}$$  \(1\)

where granulation length is denoted as $w$. Thus, the original data are divided into $(t/w)$ sub-data. Greenhouse environmental data (i.e., input) and stem runoff data (i.e., output) have different records and distinct update frequencies. Specifically, as shown in Fig. 2, stem runoff sensors update data every 10 minutes at a fixed frequency but environmental sensors update data with no fixed frequency (a. k. a. sometimes it updates data in eight minutes but sometimes it updates data in 5 minutes). In this situation, environmental data and stem runoff data are unsynchronized, which make it infeasible to perform any machine learning-based method. Despite this, we find every record of environmental data and stem runoff data has a time attribute (i.e., timestamp), and both fixed frequency and non-fixed frequency data are recorded within one hour. Thus, an intuitive idea is to transform several data records in an hour into one data record. This manipulation can keep environmental data and stem runoff data synchronized and can diminish the overall data volume. Motivated by these observations, we choose the number of data records in one hour as granulation length. Meanwhile, this agricultural IoT dataset has large volume data, hence this hour calibration granular technique not only makes environmental data and stem runoff data keep synchronized but also extracts representative information, which has the potentials for reducing computational costs. As a result, these two categories of data are transformed into the same total data records with one-by-one corresponded by hour calibration, which lays a foundation for performing data-driven models.

After the original data are divided into small sub-data, the key is how to granulate the sub-data into granules. People tend to use low, mid, and high when facing uncertain and vague situations [29]. Fuzzy membership function has many advantages when facing uncertainties and tackling tremendous data [30]. First, generally, triangular membership function is one of the most encountered membership functions in practice. Of highly applied membership functions, the triangular membership functions are formed using straight lines. These straight-line membership functions have the advantage of simplicity [31]. Second, the fuzzy triangular membership granular method could consider the comprehensive information of original data to formulate granules [32], thus continuous and voluminous data could be transformed into representative small scale data. Therefore, we deploy fuzzy granulation to break down the sub-data into three granules: $\text{Low}'_b$, $\text{Mid}'_b$, $\text{High}'_b$.

For fuzzy granulation methods, different memberships make different granules. Triangular fuzzy members with corresponding minimums, medians, and maximums are a basic but effective way to represent the comprehensive information of consecutive data. First, we order the values in each sub-data from $\{x'_j_{i+j=1}\}, j = 1, 2, \cdots, [t/w]$ into $\{x'_j_{i+j=1}\}, j = 1, 2, \cdots, [t/w]$, from small to large. Next, to determine the granules $\text{Low}'_b$, $\text{Mid}'_b$, $\text{High}'_b$, we partition the ordered sub-data $\{x'_j_{i+j=1}\}, j = 1, 2, \cdots, [t/w]$ into $\{x'_{i+j=1} + x'_{i+j=2} + \cdots, x'_{i+j=2b}\}$ and $\{x'_{i+j=2b+1} + x'_{i+j=2b+2} + \cdots, x'_{i+j=2b+2b}\}$, where $b = 1$ if $w$ is even and $b = 2$ if $w$ is odd. After that, the $\text{Low}'_b$, $\text{Mid}'_b$, $\text{High}'_b$ granules using triangular membership function can be obtained as:

$$\text{Low}'_b = \begin{cases} x'_{i+j=1} + x'_{i+j=2} + \cdots, x'_{i+j=2b} \\ [w/2] \end{cases}, j = 1, 2, \cdots, [t/w]$$

$$\text{Mid}'_b = \begin{cases} \text{Median}\{x'_{i+j=1}, x'_{i+j=2}, \cdots, x'_{i+j=2b}\} \\ [t/w] \end{cases}$$

$$\text{High}'_b = \begin{cases} x'_{i+j=2b+1} + x'_{i+j=2b+2} + \cdots, x'_{i+j=2b+2b} \\ w-[w/2]-b+1 \end{cases}, j = 1, 2, \cdots, [t/w]$$  \(2\)

Fig. 11. Framework of proposed granular-GBDT-BO.
In summary, fuzzy triangular membership granular method considers the comprehensive information of original data to formulate granules. However, it may also be vulnerable if extremes are far from medians.

B. GBDT-BO for Stem Runoff Prediction

Processed by fuzzy triangular granulation and time calibration techniques, massive multi-sensors data are broken into synchronized sub-data (i.e., Low, Mid, and High granular). Then, we treat granular environmental data as input and granular stem runoff data as output to fit GBDT model to predict cherry stem runoff, where BO aims for optimizing the hyper-parameters in GBDT. As a result, three prediction models are produced. The first is Low-GBDT-BO, which is trained by Low granular environmental and Low granular stem runoff data. The second is Mid-GBDT-BO, which is trained by Mid granular environmental, and the last is High-GBDT-BO, which is trained by High granular environmental and High granular stem runoff data.

1) Brief Recall on GBDT

GBDT comprises basic learners (i.e., decision trees) and ensemble strategies (i.e., boosting). It is assumed that \( x \) represents the set of features, including environmental data features in the work. \( F(x) \) is an approximated function of response variable \( y \) (i.e., cherry stem runoff) based on \( x \). To measure the gap between truth value and predicted value, the squared error function is utilized as a loss function:

\[
L(y, F(x)) = (y - F(x))^2
\]  

Each decision tree partitions the input space into \( J \) separated regions \( R_{m1}, \ldots, R_{mj} \) and produces a constant value \( b_m \) for the region \( R_{mj} \). Thus, each decision tree has the additive formulation:

\[
h_m(x) = \sum_{j=1}^{J} b_m I(x \in R_{mj}), \text{where}
I = 1 \text{ if } x \in R_{mj}; \text{ otherwise, } I = 0
\]  

Utilizing the training data \( \{y_i, x_i\}_{i=1}^{t} \), GBDT iteratively constructs \( M \) heterogeneous individual C&RT decisions trees \( h_1(x), h_2(x), \ldots, h_m(x) \). The updating approximating function \( F_m(x) \) and gradient descent step size \( \rho_m \) is described as follows:

\[
F_m(x) = F_{m-1}(x) + \rho_m \sum_{j=1}^{J} b_m I(x \in R_{mj})
\]  

\[
\rho_m = \arg \min \rho \left( \sum_{i=1}^{t} L(y_i, F_{m-1}(x_i) + \rho \sum_{j=1}^{J} b_m I(x_i \in R_{mj})) \right)
\]  

\( b_m \) could be replaced by a separate optimal \( y_{jm} \) for each region \( R_{mj} \). Thus, Equation (5) can be inferred as:

\[
F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} y_{jm} I(x \in R_{mj})
\]  

where

\[
y_{jm} = \arg \min_{y} \sum_{i=1}^{t} L(y_i, F_{m-1}(x_i) + y)
\]

\[
= \arg \min_{y} \sum_{i=1}^{t} (y_i - y)^2
\]  

Equation (6.1) presents that each decision tree can select all or a certain part of data (i.e., the ratio of subsample \( C \)) to construct learners. However, excessive decision trees may cause overfitting. To avoid over-fitting and to improve predictive accuracy, GBDT implements a regulation process. Learning rate, also named shrinkage, is utilized for scaling the contribution of each tree model by introducing a factor of \( \xi (0 < \xi < 1) \), as shown in (8):

\[
F_m(x) = F_{m-1}(x) + \xi \sum_{j=1}^{J} y_{jm} I(x \in R_{mj})
\]  

where the smaller \( \xi \) is, the greater the shrinkage becomes. The learning rate \( \xi \) is sensitive for GBDT. Lower \( \xi \) will iterate gradually and obtain more accurate results but at the sacrifice of longer execution time. Higher \( \xi \) saves execution time but may not converge at the real extreme. GBDT averages the results of \( m \) decision trees as final predictions, thus more trees mean more precise results but with higher computational cost. Contrarily, fewer trees will not obtain satisfactory predictions.

Based on the aforementioned analysis, we find the performance of GBDT is related to the values of its main controlling parameters (i.e., the number of decision trees in GBDT, learning rate, and the ratio of subsamples). The tuning of the hyper-parameters is a mathematical optimization problem in nature. Specifically, in GBDT forecast task, the objective is to minimize the average gap between truth value and predicted value from reduced granules, which can be formulated as below:

\[
\text{Objective function} = \min \frac{1}{t} \sum_{i=1}^{t} [y_i - \gamma]^2
\]

\[
= \min \frac{1}{t} \sum_{i=1}^{t} [y_i - g(M, \xi, C[D_{1\xi}])]^2
\]

\[
= f(M, \xi, C[D_{1\xi}])
\]

\[
\text{s.t. } 0 < M \leq m
\]

\[
0 < \xi \leq 1
\]

\[
0 < C \leq 1
\]  

where \( y_i \) denotes the truth value of observation, \( \gamma \) represents the predicted values by granular-GBDT, and \( g(M, \xi, C[D_{1\xi}]) \) can be treated as the predicted function of granular-GBDT. s.t. \( 0 < M \leq m \), \( 0 < \xi \leq 1 \), \( 0 < C \leq 1 \) denotes the constraints of decision variables and \( f \) denotes the sophisticated function measuring the gap between truth value and predicted value.

However, GBDT has no definite and analytic predictive function, thus \( f(M, \xi, C[D_{1\xi}]) \) is a black-box function [6]. Furthermore, the objective function contains original data \( D_{1\xi} \), involving \( x \) and \( y \), hence it consumes plenty of resources when it meets large-scale data volumes.

Theoretically, grid search and heuristic algorithms (such as Genetic algorithm, Ant Colony Algorithm, PSO, and others) can realize the tuning (optimization) of hyper-parameters. However, grid search is treated as a violent solution (i.e., enumeration method), and heuristic algorithms are extremely rough (i.e., no connection exists between the solving algorithm and the structure of objective function and constraints).
Furthermore, both of them are general solutions for hyper-parameters optimization.

In this cherry stem runoff prediction problem, cyber-physical system data are featured by large-scale volume with small dimensions, and GBDT has no definite analytic formulation. Thus, calls for advanced optimization methods.

2) Design of Our GBDT-BO

BO is a powerful sequential optimization strategy for finding extrema when the objective function is non-convex, or expensive to evaluate, or has no access to derivative, and it has been shown to outperform other state of the art global optimization algorithms on numbers of challenging benchmarks [33]. Gaussian Process and Acquisition Function are two core steps in BO.

a. Gaussian Process

Gaussian Process Regression aims to fit the objective function \( f(M, \zeta, C|D_{t}) \). For a Gaussian Process, we define hyper-parameter \( H = [M, \zeta, C] \) and assume that \( [H_{1},...,H_{n}] \in \chi^{n} \) and vector \( \{f(H_{1}),...,f(H_{n})\}^{T} \) follow the joint multivariable normal distribution.

\[
\begin{bmatrix}
    f(H_{1}) \\
    f(H_{2}) \ldots f(H_{n})
\end{bmatrix} 
\sim N(\mu, K)
\]

where \( n \) denotes the number of iterations, \( \mu \) denotes a \( u \times 1 \) vector and \( K \) denotes a \( u \times u \) matrix.

Gaussian Process can be characterized by its mean function \( \mu = m(H) \) and its covariance function \( K = k(H_{i},H_{j}) \). For simplicity, we assume that the prior mean is the zero function \( m(H) = 0 \). The following squared exponential function is adopted as covariance \( k(H_{i},H_{j}) = \exp(-\frac{1}{2}(H_{i} - H_{j})^{T}(H_{i} - H_{j})) \).

We sample from prior observations, choose \( \{X_{i}\} \) and sample the values of the function at these indices to produce the pairs \( \{H_{i},f_{i}\} \), where \( f_{i} = f(H_{i}) \).

When a new point comes, we denote the value of the function at the next point as \( f_{n+1} = f(H_{n+1}) \). By the properties of Gaussian processes, \( f_{i} \) and \( f_{n+1} \) follow joint Gaussian distribution:

\[
\begin{bmatrix}
    f_{i} \\
    f_{n+1}
\end{bmatrix} 
\sim N(\mu, \begin{bmatrix} K & k \\ k^{T} & k(H_{n+1},H_{n+1}) \end{bmatrix})
\]

where

\[
k = \begin{bmatrix} k(H_{n+1},H_{1}) & k(H_{n+1},H_{2}) & \ldots & k(H_{n+1},H_{n}) \end{bmatrix}
\]

Ultimately, the predictive distribution can be obtained at an expression for:

\[
f_{n+1} \sim N(\mu_{n}(H_{n+1}),\sigma_{n}^{2}(H_{n+1}))
\]

where

\[
\mu_{n}(H_{n+1}) = k^{T}K^{-1}f_{n}
\]

\[
\sigma_{n}^{2}(H_{n+1}) = k(H_{n+1},H_{n+1}) - k^{T}K^{-1}k
\]

Equation (14) gives the probabilistic function of \( f_{n+1} \). The \( \mu_{n}(H_{n+1}) \) can be treated as the predicted value, whereas \( \sigma_{n}^{2}(H_{n+1}) \) stands for the confidence of the prediction. Through this, we obtain the next combinations of hyper-parameters to fit the objective function gradually. However, due to many feasible decision points exist, which point should be iterated must be identified? To overcome it, we draw the following Acquisition Function:

b. Acquisition Function

The acquisition function aims to guide the search for the next iteration. To be more specific, we want to find:

\[
M_{n+1},C_{n+1} = \arg \max_{M,C} \alpha(M,\zeta,C|D_{t})
\]

where \( \alpha(\cdot) \) denotes the symbol for general acquisition function.

In this step, Expected Improvement (EI) is adopted as Acquisition Function because it considers the probability of improvement and magnitude of the improvement that a point can potentially yield. Particularly, we aim to minimize the expected deviation from the true extreme \( f((M,\zeta,C)) \). When a new point is selected, a one-step-ahead choice is formulated as:

\[
M_{n+1},C_{n+1} = \arg \min_{M,C} EI(f_{n}(M,\zeta,C) - f((M,\zeta,C)'))|D_{t})
\]

With the density function of joint normal distribution, \( EI \) can be obtained analytically:

\[
EI(M,\zeta,C) = \begin{cases} 
    \frac{(\mu(M,\zeta,C) - f((M,\zeta,C)'))\Phi(Z) + \sigma(M,\zeta,C)\phi(Z)}{\sigma(M,\zeta,C)} & \text{if } \sigma(M,\zeta,C) > 0 \\
    0 & \text{if } \sigma(M,\zeta,C) = 0
\end{cases}
\]

where \( \phi(\cdot) \) and \( \Phi(\cdot) \) denote the probability distribution function and cumulative distribution function of the standard normal distribution, respectively.

In summary, by calculating \( EI \), the next iteration point can be found easily and quickly. For more details, see [34]. With the above Gaussian Process and Acquisition Function, the optimal learning rate, the optimal number of decision trees and the optimal ratio of subsample in GBDT can be obtained.

IV. RESULTS AND ANALYSIS

To explore the property of the proposed Granular-GBDT-BO, experiments are performed on the IoT data mentioned in Section II. The granulated results are shown in Section A. Prediction results and comparison analysis are shown in Section B, and the results of execution time analysis are shown in Section C. Consequently, how precise prediction of stem runoff will make prominent significances in real-life agriculture is elaborated in Section D.

A. Granulated results

In this section, we use fuzzy granulation (including High, Mid and Low) and time calibration techniques to break the original data into small-scale data. Among granular computing techniques, the granulation length \( \omega \) is an extremely sensible and crucial parameter. If \( \omega \) is too large, then the volume of original data would be highly compressed, which may not retain or maintain the original data trend, as well as deteriorate the statistical distribution of original data. If \( \omega \) is extremely small, then the volume of original data would not be properly reduced. Theoretically, the granulation length should be subject
to the models and the trends of original data. Given that one record exists in every 10 minutes in sensors, \( \omega = 6 \) and \( \omega = 144 \) (24 \times 6) represent that the granule size is one hour and one day, respectively. Therefore, we utilize \( \omega = 6 \) and \( \omega = 144 \) (24 \times 6) to conduct empirical experiments from the views of data model and data statistical distribution, and the results are as follows.

1. Data model and trend

Fig. 12_A Granulated air temperature (one hour).

Fig. 12_B Granulated air temperature (one day).

Fig. 13_A Granulated air humidity (one hour).

Fig. 13_B Granulated air humidity (one day).

Fig. 14_A Granulated soil temperature (one hour).

Fig. 14_B Granulated soil temperature (one day).

Fig. 15_A Granulated soil humidity (one hour).

Fig. 15_B Granulated soil humidity (one day).

Fig. 16_A Granulated CO2 concentration (one hour).
As Figs. 12_A and 12_B show, either the granulation that length $\omega = 6$ (one hour) or $\omega = 144$ (one day) could reflect and maintain the original air temperature data patterns (i.e., it is hot in summer and cold in winter).

![Fig. 16_B Granulated CO2 concentration (one day).](image)

![Fig. 17_A Granulated effective illumination radiation (one hour).](image)

![Fig. 17_B Granulated effective illumination radiation (one day).](image)

Furthermore, we could find the fluctuations of air temperature from a day level (i.e., $\omega = 144$). Compared with one-day granulation length, one-hour granulation length may perceive the agricultural environment in greater detail with more fine grit data. The synergy of data trend/model with $\omega = 6$ (one hour) or $\omega = 144$ (one day) is also reflected in Figs. 13-17, that is, whether granulation length $\omega$ is equal to 6 or 144 could retain the original trend/model of original agricultural IoT data. The synergy of data trend/model with $\omega = 6$ (one hour) or $\omega = 144$ (one day) is also reflected in Figs. 13-17, that is, whether granulation length $\omega$ is equal to 6 or 144 could retain the original trend/model of original agricultural IoT data.

2 Data statistical distribution

Apart from the trend/model of original data, we also measure the statistical distributions of original data and granulated data with different granulation lengths (i.e., $\omega = 6$ and $\omega = 144$ (24×6)). Air temperature is selected to test the statistical distribution, and the results are as follows. Fig. 18 (a)(c)(e) indicates the distribution of original data versus the data handled by one-hour granulation method, including Low, Mid, and High granule. Analogously, Figs. 18 (b)(d)(f) indicate the distribution of original data versus the data handled by one-day granulation method, including Low, Mid, and High granule.

![Fig. 18 (a)(c)(e).](image)

![Fig. 18 (b)(d)(f).](image)

![Fig. 18 (c).](image)

Obviously, the statistical distribution of original data is in line with the statistical distribution when granulation length $\omega = 6$, as shown in Fig. 18 (a)(c)(e). However, when the granulation length $\omega$ is equal to 144 (24×6), the statistical data distribution of mid granule becomes right-skewed in contrast with the original distribution, as shown in Fig. 18 (d), and the data statistical distribution of low and high granule become left-skewed in contrast with the original distribution, as shown in Fig. 18 (b)(f). In conclusion, even though granulation length $\omega = 6$ (one hour) or $\omega = 144$ (one day) could retain the original trend/model of original agricultural IoT data, granulation length $\omega = 6$ (one hour) could retain the original data statistical distribution more efficiently, as well as retain the original trend/model of original agricultural IoT data.
From the results, we can have the following observations. First, the granulated data could relatively well retain the regular patterns (Figs. 12-17) and statistical distribution characteristics (Fig. 18) of original agricultural cyber-physical system data. Second, through time calibration granular technique, the data volume diminishes greatly with synchronized one-by-one responded data. This manipulation mitigates computing pressure as well as meets the prerequisites of performing machine learning models.

B. Prediction Results and Comparison Analysis

After obtaining small volume and one-by-one synchronized data, model training and testing could be performed. We treat environmental data as inputs and cherry runoff data as outputs to achieve predictions. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are incorporated to measure the predictive accuracy of algorithms in the testing dataset:

\[
MAE = \frac{1}{t} \sum_{i} |y_i - \hat{y}_i| \quad MSE = \frac{1}{t} \sum_{i} (y_i - \hat{y}_i)^2
\]

where \( t \) denotes the number of testing samples, \( y_i \) denotes the truth stem runoff, \( \hat{y}_i \) represents the predicted stem runoff and \( i \) stands for each data record in testing dataset. Therefore, lower MAE and MSE mean higher predictive accuracy of predictive model. To prevent overfitting, a \( k \)-fold cross-validation technique is implemented. Moreover, in order to explore the scalability and robustness of our proposed granular-GBDT-BO, we update ‘\( k \)’ from 5 to 20 with 5 intervals (i.e., 5, 10, 15, 20 folds) to imitate the change of training sample sizes and then average the values in each iteration as final results. Figs. 19-20 outline the results of MAE and MSE across Low, Mid, and High granulation methods by granular-GBDT-BO with \( k \)-fold cross-validation. Table III summarizes the optimal hyper-parameters.

![Fig. 19 Predicted MAE of cherry stem runoff by granular-GBDT-BO.](image)

**Fig. 19 Predicted MAE of cherry stem runoff by granular-GBDT-BO.**

**Fig. 20 Predicted MSE of cherry stem runoff by granular-GBDT-BO.**

**Table III. Optimal Hyper-Parameters in Granular-GBDT-BO.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Optimal Trees</th>
<th>Optimal Learning Rate</th>
<th>Optimal Sub-Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-GBDT-BO</td>
<td>191</td>
<td>0.15</td>
<td>0.763</td>
</tr>
<tr>
<td>Mid-GBDT-BO</td>
<td>191</td>
<td>0.15</td>
<td>0.526</td>
</tr>
<tr>
<td>High-GBDT-BO</td>
<td>181</td>
<td>0.15</td>
<td>0.479</td>
</tr>
</tbody>
</table>

The predicted MAEs are lower than 0.038 and predicted MSEs are lower than 0.0035 in each iteration, which indicates that predicted stem runoff nearly approximates truth stem runoff with small average errors; that is, granular-GBDT-BO performs well in terms of predictive accuracy. With the accretion of iterations, MAEs and MSEs become smaller. More iterations mean more training samples and fewer testing
samples, thus the errors decline gradually.

Moreover, in order to further empirically demonstrate the superiors of our granular-GBDT-BO in contrast with baseline machine learning algorithms, comparative experiments with the implementation of granular-support vector regression (SVR with radial basis function kernel), granular-deep belief network (DBN), granular-GBDT, and granular-GBDT-BO with $k$-fold cross-validations with iterations are performed, measured by Mean Absolute Error (MAE) and Mean Squared Error (MSE). “granular-SVR” contains “Low-SVR,” “Mid-SVR,” and “High-SVR.” Analogously, “granular-DBN” comprises “Low-DBN,” “Mid-DBN,” and “High-DBN.” Lower MAE and MSE indicate higher performances of an algorithm. Figs. 21-26 illustrates the predictive MAEs and MSEs for each granule across iterations.

Granular-GBDT-BO outperforms granular-GBDT, granular-SVM, and granular-DBN in each iteration and each granule. When compared with granular-GBDT, the learning strategy of granular-GBDT-BO is enhanced by BO. Consequently, granular-GBDT-BO is superior over granular-GBDT in terms of predictive accuracy. When compared with granular-SVR and granular-DBN, granular-GBDT could effectively extract data samples and features with the bootstrap strategy to train many decision trees, which could capture the complex relationships in data, especially when the dataset has few features, thus granular-GBDT exceeds granular-SVR and granular-DBN.

Moreover, with the increase of iterations, MAEs and MSEs become smaller. The reason is that more iterations mean more training samples and fewer testing samples, thus the errors decline gradually.
C. Execution Time Analysis

In this section, we summarize the training time of our granular-GBDT-BO, conventional GBDT, and granular-GBDT, as shown in TABLES IV-V.

**TABLE IV.**
**COMPARISON OF OUR GRANULAR GBDT-BO WITH CONVENTIONAL GBDT (METRICED BY MAE).**

<table>
<thead>
<tr>
<th>Iterations</th>
<th>Execution time of conventional GBDT(s)</th>
<th>Granules</th>
<th>Execution time of granular-GBDT (s)</th>
<th>Execution time of granular-GBD-BO (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td>Low</td>
<td>76</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>184</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>86</td>
<td>200</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Low</td>
<td>157</td>
<td>542</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>290</td>
<td>541</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>171</td>
<td>391</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Low</td>
<td>253</td>
<td>657</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>445</td>
<td>864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>301</td>
<td>622</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>Low</td>
<td>356</td>
<td>932</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>577</td>
<td>1,188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>380</td>
<td>834</td>
</tr>
</tbody>
</table>

Granular-GBDT-BO has greatly reduced the training time over conventional GBDT in each iteration. Granular-GBDT has the shortest training time among them. Amid the three granules, Low granule has the shortest execution time.

These results inspire us to explore the reasons behind the phenomena. When the number of iterations is controlled, GBDT handles the whole volume data, meanwhile granular-GBDT effectively handles small volume and representative data. Therefore, granular-GBDT consumes less execution time versus conventional GBDT. In contrast with granular-GBDT, granular-GBDT-BO is the integration of two algorithms (i.e., granular-GBDT and BO). Specifically, granular-GBDT-BO is equipped with more disparate learners, a more accurate learning rate, and better learning strategies. Therefore, granular-GBDT-BO is more complex than granular-GBDT, leading to longer execution time than granular-GBDT. Furthermore, the granular technique can be treated as a strategy of diminishing training time and Bayesian Optimization can be regarded as a strategy of prolonging training time. The decreased part compensates and even surpasses the increased part, thus the training time of granular-GBDT-BO has reduced greatly versus that of conventional GBDT.

**TABLE V.**
**THE COMPARISON OF OUR GRANULAR GBDT-BO WITH CONVENTIONAL GBDT (METRICED BY MSE).**

<table>
<thead>
<tr>
<th>Iterations</th>
<th>Execution time of conventional GBDT(s)</th>
<th>Granules</th>
<th>Execution time of granular-GBDT (s)</th>
<th>Execution time of granular-GBD-BO (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td>Low</td>
<td>73</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>184</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>76</td>
<td>207</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Low</td>
<td>188</td>
<td>458</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>290</td>
<td>507</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>183</td>
<td>414</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Low</td>
<td>303</td>
<td>655</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>445</td>
<td>762</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>324</td>
<td>647</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>Low</td>
<td>386</td>
<td>856</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>577</td>
<td>1,031</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>355</td>
<td>897</td>
</tr>
</tbody>
</table>

D. Application Analysis

In this part, we discuss what the precise prediction of cherry tree stem runoff will make positive significances in precise irrigation and plant conservations.

Traditional experience-based irrigation manner in agriculture always relies on the experiences of farmers. Irrigating less water will limit plant growth whereas irrigating too much water in greenhouse not only wastes water but also provides humid and damp conditions for the generation of pests and diseases, which is the most crucial indicator affecting agro-product yields. Resultantly, pests may rob the nutrition materials from cherry fruits, leading to the degradation of fruit quality and even reduced yields. However, farmers can precisely irrigate plants according to the precisely estimated stem runoff. Higher stem runoff denotes that the plant is full, whereas lower stem runoff represents that the plant seems inactive and calls for irrigation. From the viewpoint of application, IoT-based accurate prediction of stem runoff saves water resources, as well as has potentials for plant conservations and pest control. Thereby, precise prediction of stem runoff is a sustainable manner for agricultural production and operations.

V. Conclusion and Future Work

This work proposes a hybrid machine learning approach to estimate cherry’s evapotranspiration considering the characters of agricultural greenhouse cyber-physical systems (i.e., unsynchronized and tremendous data) by granular-GBDT-BO. The time calibration granular technique aims to break large-scale and unsynchronized data into small-scale and synchronized data with representativeness. GBDT-BO aims to achieve precise prediction for evapotranspiration estimation. IoT data from Xi’an Modern Fruit Industry Center in Shaanxi province, China verify the superiority of our proposed method.
with reduced computational cost, satisfactory and robust predictive accuracy. Resultantly, precise evapotranspiration estimation could provide operational guidance for plant irrigation, plant conservations, and pest control in the agricultural greenhouse.

However, some extensions of the work can be done in the future. In this work, we precisely predict the stem runoff of cherry trees, but when and how to irrigate plants according to the predicted stem runoff could be explored further. In terms of methodology, more recent and state-of-the-art granulation techniques could be introduced for dealing with more complicated agricultural IoT data. In addition, this work handles offline voluminous data without considering the real-time (online) data situations. So here calls for further research, which may comprise incremental learning, etc.

REFERENCES


