sensors

MDPI

Article

# Bayesian Sigmoid-Type Time Series Forecasting with Missing Data for Greenhouse Crops

**Alexander Kocian [1,2,*], Giulia Carmassi [2], Fatjon Cela [2], Luca Incrocci [2], Paolo Milazzo [1] and Stefano Chessa [1]**

[1] Department of Computer Science, University of Pisa, 56127 Pisa, Italy; paolo.milazzo@unipi.it (P.M.); stefano.chessa@di.unipi.it (S.C.)
[2] Department of Agriculture, Food and Environment, University of Pisa, 56124 Pisa, Italy; giulia.carmassi@unipi.it (G.C.); fatjon.cela@agr.unipi.it (F.C.); Luca.incrocci@unipi.it (L.I.)
* Correspondence: kocian@di.unipi.it

Received: 23 April 2020; Accepted: 4 June 2020; Published: 7 June 2020


**Abstract:** This paper follows an integrated approach of Internet of Things based sensing and machine learning for crop growth prediction in agriculture. A Dynamic Bayesian Network (DBN) relates crop growth associated measurement data to environmental control data via hidden states. The measurement data, having (non-linear) sigmoid-type dynamics, are instances of the two classes observed and missing, respectively. Considering that the time series of the logistic sigmoid function is the solution to a reciprocal linear dynamic model, the exact expectation-maximization algorithm can be applied to infer the hidden states and to learn the parameters of the model. At iterative convergence, the parameter estimates are then used to derive a predictor of the measurement data several days ahead. To evaluate the performance of the proposed DBN, we followed three cultivation cycles of micro-tomatoes (MicroTom) in a mini-greenhouse. The environmental parameters were temperature, converted into Growing Degree Days (GDD), and the solar irradiance, both at a daily granularity. The measurement data were Leaf Area Index (LAI) and Evapotranspiration (ET). Although measurement data were only available scarcely, it turned out that high quality measurement data predictions were possible up to three weeks ahead.

**Keywords:** prediction; sigmoid; time-series; dynamic Bayesian network; missing data; evapotranspiration; leaf area index; MicroTom; IoT

## 1. Introduction

The need to supply food to a growing population leads agriculture to deal with new and significant challenges. High-input and resource-intensive farming systems, which have caused massive deforestation, water scarcities, soil depletion, and high levels of greenhouse gas emissions, cannot deliver sustainable food and agricultural production. Instead, agriculture will increasingly need innovative systems that protect and enhance the natural resource base, while increasing productivity [1]. Precision agriculture has a significant potential to reduce agricultural inputs, enhance agricultural sustainability, and increase production, as it allows greater accuracy in targeting the correct amount of inputs, at the correct time, and in the correct location compared to conventional agricultural methods [2]. Within the precision agriculture approach are currently also included Internet of Things (IoT) technologies [3]. In a recent review [4], the authors analyzed the existing literature on the concept of IoT applications in precision agriculture, and they concluded that IoT may contribute significantly to modern agriculture, by providing automated and even remote control of farms, thus enabling a more effective management of agricultural activities, in this innovation supported by new age farmers that have already moved away from traditional farming to engage with technological farming.

However, with the introduction of IoT comes the problem of interpreting the large amount of data that are produced with this technology, which passes through to the need to develop new models capable of processing capillary and high resolution data, a requirement that can be met by the massive introduction of data-driven Artificial Intelligence (AI) techniques [5–7]. IoT technologies are also suitable for use in the greenhouse production systems where the environment is controlled, often equipped with sensors to monitor environmental and crop parameters, and the data can be acquired and processed [8]. Recent advancements in IoT, AI, and Information and Communication Technologies (ICT) in general have the potential to address some of the environmental, economic, and technical challenges, as well as opportunities in this sector [9]. Protected cultivation has rapidly expanded in many regions all over the world [10]. This cultivation system is constantly evolving; progress is continuous; and the environmental concerns are increasingly important in the greenhouse production system. New approaches and technologies are oriented toward reducing any environmental impact and an optimal management of the input. Real-time estimation of the input and the combining of a monitoring approach and modeling are growing in importance [10]. In a recent review on the impact of protected vegetable cultivation [11], the importance of good agricultural practices, especially with respect to irrigation or fertilization, was reported for the reduction of greenhouse gas emissions.

The applications of AI technologies in agriculture are many and cover several different aspects, addressing crop, water, and soil management. Concerning crop management, the applications span from yield prediction to the recognition and classification of diseases, weed detection, crop quality, species recognition, etc. [5]. A commonly used technique in agriculture is Support Vector Regression (SVR) [12,13]. Equipped with Radial Basis Functions (RBFs) as the kernel, the SVR in [14] was capable of predicting non-linear rice yield after training with historical data covering a period of 20 years. Furthermore, back-propagation and RBF neural networks have been successfully applied to (non-linear) price forecasting in agriculture after a training period of 365 days [15]. Note that the size of the training set needs to be at least one magnitude larger than that of the feature space. Among several AI methods, the Bayesian approach is particularly appealing due to its ability to make accurate predictions even with limited datasets. Here, a statistical model for the crop growth turns a first guess of the state-variable distributions, i.e., the prior distribution, along with measurements into a posterior distribution. The parameters of the model are assumed to be known. As a matter of a fact, Bayesian models have been successfully used to predict fruit yield [16] and disease development [17,18]. As concerns this work, we are interested in AI models suitable for predicting the growth of crops that may allow farmers to better plan their control actions, optimize the input, and quantitatively evaluate (through simulations) the effects of alternative interventions when the model parameters are unknown. A significant work on how historical growth, soil characteristics, and environmental parameters can be used to predict the time series of crop growth-related indicators such as Leaf Area Index (LAI) and evapotranspiration rate was given in [19]. However, the self-learning model only supported crops with exponential growth dynamics. Furthermore, the measurement data must be complete.

Nevertheless, some of the measurement data may be missing due to sensor failure, the battery's low charge, or the measurement process (for example, destructive samples are only available in a limited quantity). Missing data randomness can be divided into three classes: (i) Missing At Random (MAR): The probability that an observation is missing only depends on the observed data, but is independent of the missing data. When the model parameters are independent of the parameters of the missing data, the missing data mechanism is ignorable [20]. (ii) Missing Completely At Random (MCAR), a special case of MAR, where the probability that an observation is missing is independent of the observed and missing data. (iii) Not missing a random is data that are neither MAR nor MCAR.

Most data analysis techniques in the literature, including SVR and neural networks, have no native mechanism that handles missing data and, hence, exclude observations with missing variable values, introducing a bias. Another strategy is simple imputation, in which each missing value is substituted [21]. For example, each missing value can be imputed by interpolation techniques (for example, linear, quadratic, cubic, and nearest neighbor interpolation) [22]. Following this approach,

missing values are treated as they were known in the analysis. Consequently, the bias is reduced. Multiple imputation, in contrast, samples multiple values from the probability distribution of observed sensor data and takes the statistical mean, variance, and confidence interval to fill the gap [20]. A standard approach for multiple imputation is the Data Augmentation (DA) algorithm, which is a Markov chain Monte Carlo technique [23]. In Bayesian inference, the missing data can be resolved by iterating between the imputation step and posterior step [21]. The most efficient method, we follow subsequently, is to approach the maximum likelihood estimate in the presence of missing data using the expectation-maximization algorithm [24]. In contrast to the stochastic approximation in [25], our solution is exact.

In this study, we develop a DBN for time series forecasting with daily granularity of crop growth in the IoT based mini-greenhouse COLTIVazione Automatizzata Miniaturizzata Innovativa (COLTIV@MI) [26]. Our DBN learns the model parameters ad-hoc without the need for time consuming training cultivation cycles. Our DBN relates the non-linear crop growth associated measurement data LAI and ET to the environment associated control data GDD and solar irradiance via hidden states. Many plants have a logistic growth evolution. To resemble such growth dynamics, the DBN will be based on a highly complex non-linear model. We will show that the reciprocal function, though, follows a linear dynamic model with low computational complexity. Often, measurement data are only available scarcely at an irregular time spacing. For example, the LAI has been obtained by destructive measurement once every three weeks. To learn the parameters of the model in Gaussian noise when measurement data are MAR, we derive a novel tracking algorithm based on the exact Expectation-Maximization (EM) framework. The self-trained model is then used to derive a measurement predictor many days ahead. As an example, we consider the micro-tomatoes MicroTom [27] through the proposed approach, which can be applied to any other crop type with sigmoid-type growth dynamics. We will see that LAI and ET can be predicted up to 21 days ahead with high accuracy even if almost all of the measurement data are missing.

The paper is organized as follows. Section 2 develops the DBN, the underlying system model, and the EM algorithm to learn the parameters of the model. The plant growth parameter estimates at iterative convergence are then used to derive a multiple-step ahead predictor. The experimental settings will be discussed. The experimental results are discussed in Section 3 followed by conclusions and limitations in Section 4.

## 2. Materials and Methods

### *2.1. The Standard Models and Beyond*

The LAI is a dimensionless measure for the total area of leaves per unit projected ground area and directly related to the amount of light that can be intercepted by plants. The LAI is a key parameter to predict photosynthetic biomass production. For tomatoes, it was shown in [28] that LAI from planting to the time of harvest can be modeled as a function of thermal time, expressed in growing degree days (GDD) (°C). The resulting function:

$$\mathrm{LAI}_t = \alpha + \frac{\beta - \alpha}{1 + \exp\left\{\left(\zeta - \sum_{t' \le t} \mathrm{GDD}_{t'}\right)/\delta\right\}}. \tag{1}$$

has the form of a Boltzmann sigmoid where $\alpha$, $\beta$, $\zeta$, and $\delta$ are constants to be obtained by regression analysis. The index $t$ in (1) determines the $t^{\mathrm{th}}$ Growing Day after Transplantation (GDT). For the calculation of GDD from average daily temperature, a value of 10 °C was selected as a base temperature according to [29].

The Evapotranspiration (ET) is a combination of the water transpired by plants during the growth or retained in the plant tissue plus the moisture evaporated from the soil surface and vegetation. Several ET models have been developed during the last few years, all based on the Penman–Monteith approach [30], which is a worldwide accepted modeling approach to determine evapotranspiration.

The current use of the Penman–Monteith model is based on the calculation of ET for outdoor climates, while the Stanghellini ET model [31] was implemented in high technology controlled environment greenhouses. Both equations require several parameters related to the weather, as well as the crop stomatal and aerodynamic resistances, which are not always available. Alternative relatively simple empirical models have been developed for irrigation scheduling purposes, while the other knowledge based mechanistic models have been developed for climate control purposes [32]. These models take into account greenhouse climate variables such as radiation and vapor pressure deficit and crop measurements such as the Leaf Area Index (LAI) or stomatal resistance [28,33–39]. In some cases, a multiple linear regression of ET against vapor pressure deficit, as well as outside or inside solar radiation has been proposed for irrigation management in greenhouse crops [40]. For tomatoes, it was shown in [28] that the Baille model in [33] simplifies to:

$$\mathrm{ET}_t = \epsilon(1 - \exp\{-k\,\mathrm{LAI}_t\})\frac{R_t}{\lambda} + \gamma. \tag{2}$$

The ET is expressed in mm $\mathrm{d}^{-1}$; $R_t$ is the daily value of solar irradiance (MJ $\mathrm{m}^{-2}$ $\mathrm{d}^{-1}$); $\lambda$ is the latent heat of water vaporization (2.45 MJ $\mathrm{kg}^{-2}$); $k$ is the light extinction coefficient of the canopy (measured as 0.69); and $\epsilon$ and $\gamma$ are the regression parameters. Inspecting (2), it can be seen that the ET is a modulated sigmoid function with a negative exponential envelope. The models (1) and (2) act as the benchmark for our dynamic Bayesian growth model.

Difference equations can be used to predict the growth status one discrete time step ahead. Specifically, the time series of the LAI in (1) resembles the scaled and time-shifted logistic function $z_t = [\kappa(1 + \exp\{-\mu(t - t_0)\})]^{-1}$ with the system parameters $\kappa$, $\mu$, and $t_0$ denoting the steady state, decay, and the point of inflection, respectively. The logistic function is the solution to a non-linear difference equation, which is cumbersome to implement. Its reciprocal dynamics,

$$x_t \triangleq z_t^{-1} = \kappa\left(1 + \exp\{-\mu(t - t_0)\}\right), \tag{3}$$

however, has an exponential shape that can easily be generated by the first-order linear ordinary difference equation:

$$\frac{x_{t+1} - x_t}{\Delta} = -\mu\, x_t + \kappa\mu \tag{4}$$

at low computational complexity. Here, $\Delta$ denotes the time granularity. Note that the first term on the right-hand side of (4) depends on the latent variable, while the second does not. On the other hand, the ET time series in (2) has the form:

$$x_t = \kappa'\left(1 - \exp\{-\mu' t\}\right). \tag{5}$$

Note that despite the different sign in front of the exponential, the expressions in (3) and (5) are the solution to the same difference equation in (4), but with distinct initial conditions.

### *2.2. Dynamic Bayesian Network*

We now derive a Dynamic Bayesian Network (DBN) for stochastic crop growth with the sigmoid-type activation function when observation data are sparse. When the underlying process is Markovian, the DBN replicates a particular template over discrete steps in time. The template is a directed acyclic graph, representing the state transition distribution from one state to the next state and the emission distribution within the same state. The edges of the DBN reflect a conditional dependency, while the nodes correspond to one of the three kinds of variables. The arrows indicate the conditional dependencies. The expression in (4) motivates us to model the reciprocal dynamics of crop growth as DBN, having a length of $T$ growing days. We subsequently define the states, derive the template, and draw the DBN.

For the states, the transition probability distribution from hidden growth state $\{\boldsymbol{x}_t : \boldsymbol{x}_t \in \mathbb{R}^K, t \in [1, T]\}$ to state $\boldsymbol{x}_{t+1}$,

$$p\left(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t+1}; (\boldsymbol{I} - \boldsymbol{A}\,\boldsymbol{U}_t)\boldsymbol{x}_t + \boldsymbol{B}\,\boldsymbol{u}_t, \boldsymbol{\Sigma}_n\right), \tag{6}$$

follows a first-order Markov process with additional Gaussian noise. Here, $\mathcal{N}(\cdot; \boldsymbol{m}, \boldsymbol{\Sigma})$ denotes a normal distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$. Moreover, $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ denote the state matrix and the control matrix, respectively; $\boldsymbol{I}$ is the $K$-dimensional identity matrix, and the diagonal matrix $\boldsymbol{U}_t \triangleq \operatorname{diag}\{\boldsymbol{u}_t\}$. The control vector $\boldsymbol{u}_t$, $\boldsymbol{u}_t \in \mathbb{R}^K$ is a deterministic function of the environmental parameters independent of the crop.

The template of the DBN is comprised of the above state distribution and subsequent emission distribution:

$$p\left(\boldsymbol{y}_t^{(\text{obs})}|\boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{y}_t^{(\text{obs})}; \boldsymbol{C}\boldsymbol{x}_t, \boldsymbol{\Sigma}_w\right) \tag{7}$$

instead of having perfect knowledge of the state variable. Here, the measurement matrix is denoted by $\boldsymbol{C} \in \mathbb{R}^{1\times K}$. Moreover, $\boldsymbol{y}_t^{(\text{obs})} \in \mathbb{R}$ denotes the observed data. The remaining measurement data $\boldsymbol{y}_t^{(\text{mis})} \in \mathbb{R}$ are conceptual and denote the data that were not observed.

For the initial state,

$$p\left(\boldsymbol{x}_1\right) = \mathcal{N}\left(\boldsymbol{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right). \tag{8}$$

With the state transition model defined in (6) and the emission model in (7), we are now ready to derive the DBN as illustrated by two time slices in Figure 1.

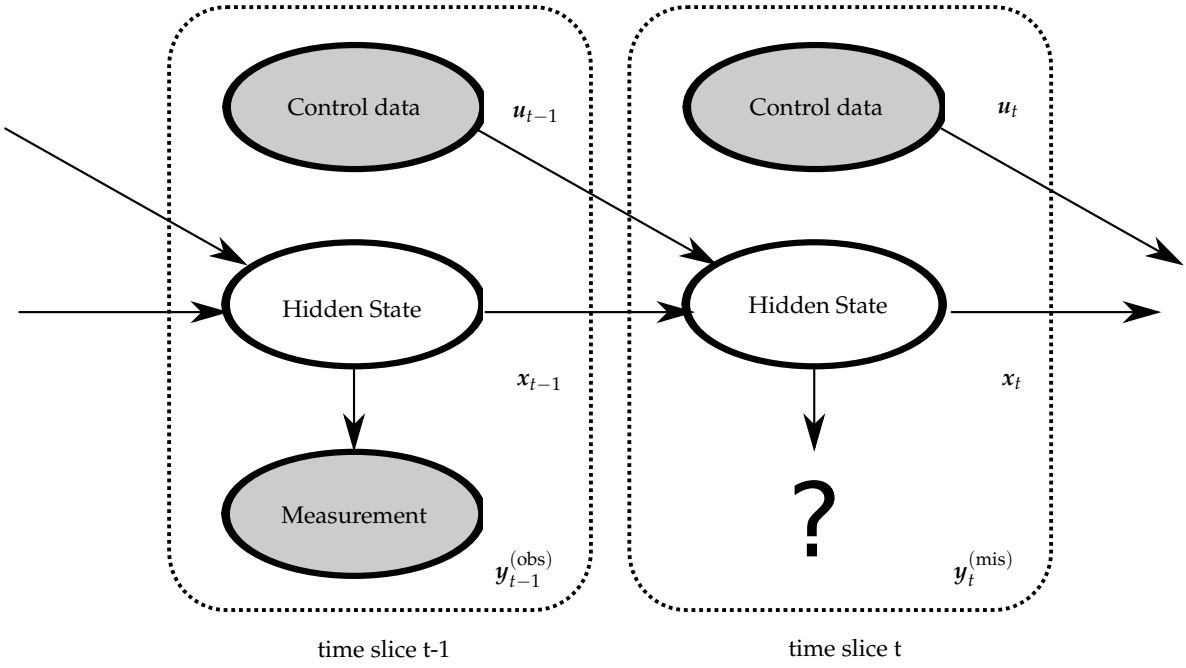


**Figure 1.** Two time slice Bayesian network with missing measurement data.

In the sequel, the parameter vector $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Sigma}_n, \boldsymbol{C}, \boldsymbol{\Sigma}_w, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}$, the latent state sequence $\boldsymbol{x} \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, and the missing data are unknown and, hence, require estimation.

#### 2.2.1. The EM Algorithm

This section derives a maximum-likelihood based tracking algorithm for the state sequence. So far, we have designed the DBN in Figure 1. Unrolling the DBN for $T$ time slices, it follows for the joint state-measurement probability distribution that:

$$p\left(\boldsymbol{x}, \boldsymbol{y}^{(\text{obs})}, \boldsymbol{y}^{(\text{mis})} | \boldsymbol{\theta}\right) = p\left(\boldsymbol{x}_1 | \boldsymbol{\theta}\right) p\left(\boldsymbol{y}_1^{(\text{obs})}, \boldsymbol{y}_1^{(\text{mis})} | \boldsymbol{x}_1, \boldsymbol{\theta}\right) \prod_{t=2}^{T} p\left(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right) p\left(\boldsymbol{y}_t^{(\text{obs})}, \boldsymbol{y}_t^{(\text{mis})} | \boldsymbol{x}_t, \boldsymbol{\theta}\right). \quad (9)$$

Due to the Markov property in (6), the joint state-measurement distribution decomposes into the product of individual conditional distributions. We now have two options for handling missing observations: (i) modeling the missing data mechanism or (ii) ignoring it, which is equivalent to integrating out the missing observations from the joint density function in (9). Following this approach, we get:

$$p\left(\boldsymbol{x}, \boldsymbol{y}^{(\text{obs})} | \boldsymbol{\theta}\right) = \int p\left(\boldsymbol{x}, \boldsymbol{y}^{(\text{obs})}, \boldsymbol{y}^{(\text{mis})} | \boldsymbol{\theta}\right) d\boldsymbol{y}^{(\text{mis})}. \quad (10)$$

Rubin showed in [20] that sufficient conditions for ignoring missing data are:

- The data are MAR, i.e., the missing data mechanism is only allowed to depend on $\boldsymbol{y}^{(\text{obs})}$;
- the model parameters governing absence and the parameters of interest $\boldsymbol{\theta}$ reside in different spaces.

In our case, both conditions are fulfilled, so that we may safely ignore the missing data mechanism.

Based on the marginal distribution in (10), we now use the EM algorithm to approximate the Bayesian inference of the state sequence. The EM algorithm postulates complete (hidden) data $\mathcal{X}$ that would ease the computation of $\boldsymbol{\theta}$ if they were known. Following this approach, let $\mathcal{X}$ contain the reciprocal dynamics of plant growth along with the observed measurement data, i.e., $\mathcal{X} = \{\boldsymbol{x}, \boldsymbol{y}^{(\text{obs})}\}$. Starting from iteration $i = 0$, the E-step of the algorithm provides the expected value of the entire complete data given the observed measurements and a guess of the parameter vector, i.e.,

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{[i]}) = \mathbb{E}\{\ln p\left(\boldsymbol{x}, \boldsymbol{y}^{(\text{obs})} | \boldsymbol{\theta}\right) | \boldsymbol{y}^{(\text{obs})}, \boldsymbol{\theta}^{[i]}\}. \quad (11)$$

The E-step iterates between prediction and correction if measurement data are available or only predicts the complete data based on previous state information without a response when measurement data are missing according to (10). The M-step,

$$\boldsymbol{\theta}^{[i+1]} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{[i]}), \quad (12)$$

is learning from the updated complete data in the E-step. The sequence of log-likelihood values $\left\{\ln p\left(\boldsymbol{y}^{(\text{obs})} | \boldsymbol{\theta}^{[i]}\right)\right\}_{i=0}^{\infty}$ is non-decreasing and converges to a stationary point [24,41].

At iterative convergence, the expected state sequence has the form:

$$\boldsymbol{x}_t^{[\infty]} = \mathbb{E}\{\boldsymbol{x}_t | \boldsymbol{y}^{(\text{obs})}, \boldsymbol{\theta}^{[\infty]}\} \quad (13)$$

with the error variance:

$$\boldsymbol{V}_t^{[\infty]} = \text{Cov}\{\boldsymbol{x}_t | \boldsymbol{y}^{(\text{obs})}, \boldsymbol{\theta}^{[\infty]}\}. \quad (14)$$

The derivation of (13) and (14) is provided in Appendix A.

#### 2.2.2. Prediction of Measurement Data

If control data were available $q$ days beyond the current growing day $T$, but measurement is missing, the expected state-sequence in (13)) would already be the solution to our problem. In practice,

however, control data are only available until time step $T$. A simple yet efficient predictor is to deploy (13), but keep control data frozen at time step $T$. Following this approach, we obtain:

$$\boldsymbol{x}_{T+q}^{[\infty]} = \left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{U}_T\right)\boldsymbol{x}_{T+q-1}^{[\infty]} + \boldsymbol{B}^{[\infty]}\boldsymbol{u}_T. \tag{15}$$

Substituting (15) for (7), it follows for the $q$-step ahead predictor of the measurement data that:

$$\boldsymbol{y}_{T+q}^{[\infty]} = \boldsymbol{C}^{[\infty]}\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{U}_T\right)\boldsymbol{x}_{T+q-1}^{[\infty]} + \boldsymbol{C}^{[\infty]}\boldsymbol{B}^{[\infty]}\boldsymbol{u}_T. \tag{16}$$

having error variance:

$$\boldsymbol{\Sigma}_{y,T+q}^{[\infty]} = \boldsymbol{C}^{[\infty]}\left(\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{U}_T\right)\boldsymbol{V}_{T-1+q}\left(\boldsymbol{I} - \boldsymbol{A}^{[\infty]}\boldsymbol{U}_T\right)^T + \boldsymbol{\Sigma}_{\boldsymbol{n}}^{[\infty]}\right)\left(\boldsymbol{C}^{[\infty]}\right)^T. \tag{17}$$

The above predictor runs freely without response.

#### 2.2.3. Initialization of the EM Algorithm

The stationary point of the log-likelihood function, reached by the EM algorithm, highly depends on the initial point [42]. Hence, initialization requires special care.

For the noise covariances, we start off with small values on their diagonal:

$$\boldsymbol{\Sigma}_{\boldsymbol{n}}^{[0]} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{[0]} = \boldsymbol{\Sigma}_1^{[0]} = \epsilon\boldsymbol{I}, \quad \epsilon \ll 1; \tag{18}$$

The measurement matrix $\boldsymbol{C}^{[0]}$ is initialized as:

$$\boldsymbol{C}^{[0]} = \boldsymbol{E} \tag{19}$$

where $\boldsymbol{E}$ is the all-one matrix. Substituting $\boldsymbol{C}^{[0]}$ for (8), the state-sequence starts off with:

$$\boldsymbol{\mu}_1^{[0]} = \left(\boldsymbol{C}^{[0]}\right)^T\left(\boldsymbol{C}^{[0]}\left(\boldsymbol{C}^{[0]}\right)^T\right)^{-1}\boldsymbol{y}_1^{(\mathrm{obs})}. \tag{20}$$

The initial matrices $\boldsymbol{A}^{[0]}$ and $\boldsymbol{B}^{[0]}$ depend on the growth parameter. For the LAI, the reciprocal logistic curve in (3) is a function of three parameters $\kappa$, $\mu$, and $t_0$, which are the solutions to three equations:

$$\begin{aligned}
\mu &= \frac{\log\left(\boldsymbol{y}_{t_2}^{(\mathrm{obs})}/\kappa - 1\right) - \log\left(\boldsymbol{y}_{t_1}^{(\mathrm{obs})}/\kappa - 1\right)}{t_1 - t_2} \\
t_0 &= \frac{\log\left(\boldsymbol{y}_{t_1}^{(\mathrm{obs})}/\kappa - 1\right) + \mu t_1}{\mu} \\
\kappa &= \frac{\boldsymbol{y}_{t_3}^{(\mathrm{obs})}}{1 + \exp\left\{-\mu(t_3 - t_0)\right\}}
\end{aligned} \tag{21}$$

This system of nonlinear equations, however, has no closed-form solution. For MicroTom tomatoes, the reciprocal LAI has the steady state of $\kappa \approx 1$ [43]. Having fixed $\kappa$, the computation of $t_0$ is obsolete, and $\mu$ has the closed-form solution in (21). From (4) and (6), it can be seen that the state matrix $\boldsymbol{A}$ is proportional to the decay. Ergo,

$$\boldsymbol{A}^{[0]} = \mu\ \mathrm{diag}\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t\}^{-1}. \tag{22}$$

We normalized $\boldsymbol{A}^{[0]}$ by the sample mean of the control signal. As the control matrix $\boldsymbol{B}$ is proportional to the decay and the steady state, it follows:

$$\boldsymbol{B}^{[0]} = \boldsymbol{A}^{[0]}\kappa. \tag{23}$$

For the ET, the negative exponential decay in (5) can be specified by two observation points according to:

$$\begin{aligned} \mu' &= -\frac{1}{t_1}\left(\log(1 - \boldsymbol{y}_{t_1}^{(\text{obs})}/\kappa')\right) \\ \kappa' &= \frac{\boldsymbol{y}_{t_2}^{(\text{obs})}}{1 - \exp\left\{-\mu' t_2\right\}} \end{aligned} \tag{24}$$

which can be solved in an iterative fashion. Substituting $\mu'$ for $\mu$ in (22), as well as $\kappa'$ for $\kappa$ in (23), we again obtain initial guesses for the matrices $\boldsymbol{A}^{[0]}$ and $\boldsymbol{B}^{[0]}$, respectively.

With this type of initialization, our EM algorithm is able to find the global maximum of the likelihood function with the aid of only two initial observations.

### *2.3. Experimental Design*

Trials were conducted at the Department of Agriculture, Food and Environment at the University of Pisa, Pisa, Italy (latitude 43°42′ N, longitude 10°24′ E). MicroTom plants were cultivated in plastic planter pots $10 \times 10 \times 17$ cm filled with perlite/peat moss at a density of 12 plants $\text{m}^{-2}$ The plants were irrigated with a nutrient solution with the following nutrient ion composition (expressed as mol $\text{m}^{-3}$): $\text{N-NO}_3^-$ 10.4, $\text{H}_2\text{PO}_4^-$ 1.0, $\text{K}^+$ 7.5, $\text{Mg}^{2+}$ 2.0, $\text{Ca}^{2+}$ 4.5, plus Hoagland concentration of trace elements. The Electrical Conductivity (EC) of the nutrient solution was 2.34 dS $\text{m}^{-1}$, and the pH was 5.5. The experiments were conducted between 1 April 2019 and 31 May 2019 in three different experimental environments. Environment 1 (Env 1) was a warm greenhouse with a high level of light; Environment 2 (Env 2) was shaded and equipped with emergency heating only; while Environment 3 (Env 3) was a first prototype of a domestic greenhouse for indoor cultivation equipped with LED lighting system. The three growth environments were chosen such that the growth of MicroTom could be evaluated at different conditions of light and temperature, which are the main parameters influencing crop development and crop evapotranspiration. The environmental parameters temperature converted into Growing Degree Days (GDD) and the solar irradiance $R$ were sensed (Netsens, Sesto Fiorentino, Tuscany, Italy) and recorded (Netsens Wireless Unit Model No. MN-0086-AE). Table 1 lists their descriptive statistics.

The different climatic conditions produced important differences in cumulated ET and growth. The ET was measured in the mornings by using a scale. Roughly 50 percent of all data points, however, were missing. Leaf area measurements were made on 5 plants as the sum of the area of the leaves of each plant by a planimeter (DT Area Meter MK2, Delta T-Devices, Cambridge, U.K.), starting from Growing Days after Transplantation (GDT) equal to zero and ending at the ultimate GDT with two samples in between.

The proposed DBN was benchmarked against the non-linear growth model by Carmassi in Section 2.1, as well as a low-complex Linear time-series Regression Model (LRM). For round-fruit tomatoes (*Solanum lycopersicum* L. cv. Jama F1), the parameters of the model were given in [28]. Since the vegetative habitus of our cv. "MicroTom" was quite different, it was necessary to re-calibrate the model parameters. Using the dataset of Env 1, it followed for the LAI reference model in (1) by Carmassi that $\alpha = 0.03$, $\beta = 0.90$, $\zeta = 560$, and $\delta = 50$ when the base temperature was 10 °C. Following the same approach for the ET reference model in (2), we obtained $\epsilon = 0.109$ and $\gamma = 0.32$ mm $\text{d}^{-1}$.

**Table 1.** Descriptive statistics of the control parameters in three different climatic environments. Env, Environment.

| | | Temperature (°C) | GDD (C°) | Irradiance $R$ (MJ $m^{-2}$ $d^{-1}$) |
|---|---|---|---|---|
| Env 1 | Mean | 22.86 | 12.86 | 12.82 |
| | Maximum | 27.29 | 17.3 | 21.18 |
| | Minimum | 18.59 | 8.6 | 2.35 |
| | Accumulation | N/A | 1158.7 | 820.67 |
| Env 2 | Mean | 18.40 | 8.4 | 4.70 |
| | Maximum | 24.57 | 14.6 | 8.20 |
| | Minimum | 12.08 | 2.1 | 1.49 |
| | Accumulation | N/A | 857.2 | 291.44 |
| Env 3 | Mean | 19.14 | 9.14 | 1.29 |
| | Maximum | 23.63 | 13.60 | 1.29 |
| | Minimum | 15.31 | 5.30 | 1.29 |
| | Accumulation | N/A | 921.1 | 82.56 |

The competing LRM used the same training data as our DBN did. Due to lack of measurement data, the LRM describes a linear relationship between LAI and cumulative GDD, only, according to:

$$\mathrm{LAI}_t = \rho_0 + \rho_1 \sum_{t' \leq t} \mathrm{GDD}_{t'} + \varepsilon_t \tag{25}$$

where $\rho_0$ and $\rho_1$ are the regression coefficients using a standard least squares fit and $\varepsilon$ is the model error. For the ET, the LRM assumes a linear relation between the dependent variable ET and the independent variables GDD accumulated over time and $R$ according to:

$$\mathrm{ET}_t = \rho_0 + \rho_1 \sum_{t' \leq t} \mathrm{GDD}_{t'} + \rho_2 R_t + \varepsilon_t. \tag{26}$$

Finally, it should be noted that the EM algorithm in our DBN was iterated until the log-likelihood function $\ln p\left(\boldsymbol{y}^{(\mathrm{obs})} | \boldsymbol{\theta}^{[i]}\right)$ increased by less than $1‰$ or the limit of 100 iterations was reached.

## 3. Results and Discussion

Given the environmental parameters in Section 2.3 and sparse historical measurement data until some time instant $T$, our DBN was deployed to predict the posterior distribution $q$ growing days ahead.

In MicroTom LAI prediction, the measurement data were observed in only 6% of all cases. The other 94 % were treated as MAR. The control data, however, were available at all time instants. The EM algorithm operated on the reciprocal LAI. The first two measurement points were used to initialize the algorithm according to Section 2.2.3 with $\kappa = 1$. Figure 2 reports the performance of our DBN as a function of GDT. At the day of measurement, our predictor forecast the next measurement, which was $q = 21$ days (very long) ahead. In the upper subplot, the blue markers indicate the reciprocal observations at $T = \{1, 22, 43, 64\}$. The black curves show the mean prediction value $y_{T+21}^{[\infty]}$ of 1/LAI in (16). Its standard deviation $\boldsymbol{\Sigma}_{T+21}^{[\infty]}$ in (17) was encoded as half the width of the filled region around the mean value. Clearly, the predictor only had access to historical measurement data until time instant $T$. The lower subplot shows the LAI vs. GDT using the same line-styles. For comparison purposes, the Carmassi model in (1) and the LRM in (25) are also added to the plot as red and brown lines, respectively. It can be seen that our DBN correctly anticipated the reflection point of the sigmoid function in all three cultivation environments despite the extremely low number of measurement points. Clearly, with the increasing number of observed data points, the quality of the LAI forecasts improved. The average prediction error, averaged over all $T$, was equal to 15.5%, 12.2%, and 19.7% for Env 1, Env 2, and Env 3, respectively. Details are listed in Table 2. Note that the inverse of the

reciprocal mean depicted in the lower subplot is a lower bound of the mean value shown in the upper subplot (Jensen's inequality). Moreover, the error variance of the reciprocal of a Gaussian random variable tends to infinity and, hence, was omitted from the lower subplot. The analytic Carmassi model accurately estimated the inflection point in Env 3 excluding the reference environment Env 1. The simple LRM in (25) could be seen as first-order approximation of the former in (1) and, hence, was quite accurate in the neighborhood of the inflection point of the sigmoid function, but became more erroneous the larger the time deviation was.

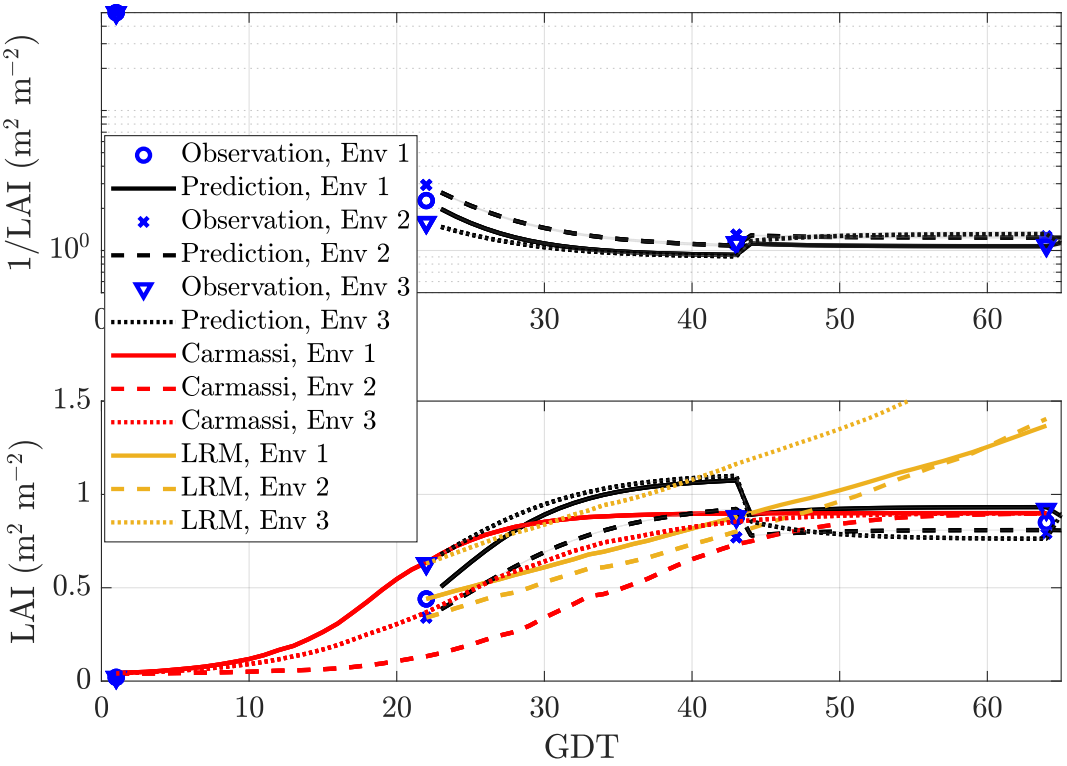


**Figure 2.** Prediction of the LAI with prediction length $q = 21$ for micro-tomatoes (MicroTom) with sparse data in three environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination). LRM, Linear time-series Regression Model.

**Table 2.** Prediction error of LAI vs. GDT for MicroTom with sparse data in three environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination).

| **Error (%)** | $q = 1$ | | $q = 21$ | |
|---|---|---|---|---|
| | $T = 43$ | $T = 64$ | $T = 43$ | $T = 64$ |
| DBN Env 1 | 16.86 | 13.03 | 22.48 | 9.17 |
| DBN Env 2 | 25.64 | 6.36 | 21.30 | 2.74 |
| DBN Env 3 | 8.34 | 9.68 | 23.74 | 14.82 |

Figure 3 shows the performance of our DBN with prediction length $q = 1$ day (very short). At time $T$, the EM algorithm either observed or missed the measurement point, requiring reconstruction. The resulting curve, influenced by the different control data every day, was rougher, but generally speaking, more accurate than for a prediction length of $q = 21$. In particular for Env 3, the prediction error was now under 10%.

Table 3 indicates that the mean estimation error for the analytical Carmassi model was equal to 19.42% (13.46%) for Env 2 (Env 3). This error was dominated by the loose approximation near the initial growth state. The mean estimation error of the LRM was equal to 23.88 %, 34.0%, and 50.0% for

Env 1, Env 2, and Env 3, respectively. In contrast to the Carmassi model, this error was dominated by the steady growth state farthest away from the likely missed inclination point due to the nature of the first order approximation.

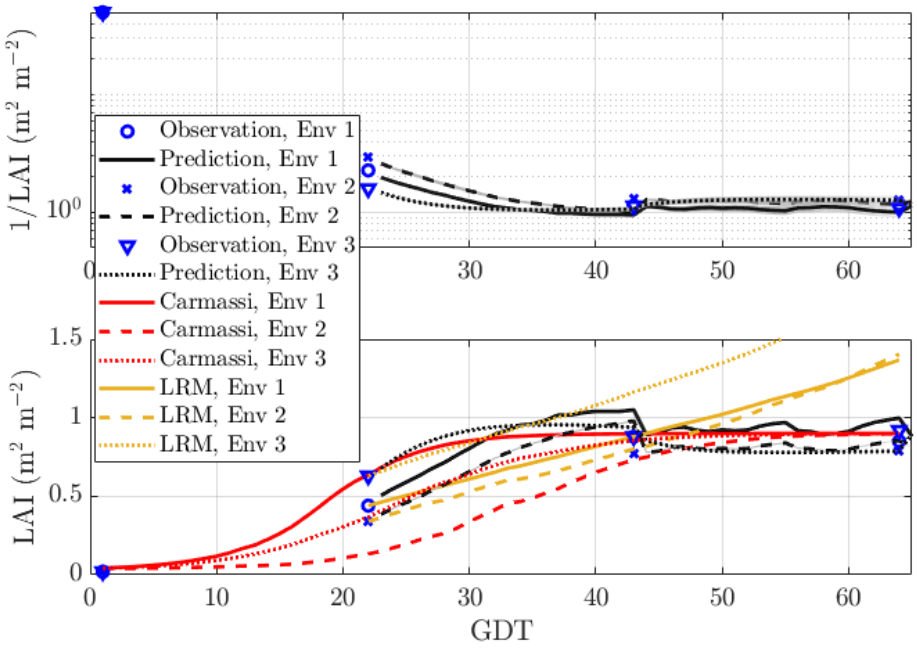


**Figure 3.** Prediction of the LAI with prediction length $q = 1$ for MicroTom with sparse data in three environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination).

**Table 3.** Prediction error of the Carmassi and the Linear Regression Model (LRM) for LAI vs. GDT. The parameters of the former were adjusted during the cultivation cycle in Env 1, while the slope and intercept of the latter during the first two measurement points in each cultivation cycle.

| **Error (%)** | $T = 1$ | $T = 22$ | $T = 43$ | $T = 64$ |
|---|---|---|---|---|
| Carmassi Env 2 | 102.76 | 55.61 | 16.44 | 13.92 |
| Carmassi Env 3 | 109.19 | 41.55 | 2.90 | 2.24 |
| LRM Env 1 | N/A | N/A | 0 | 61.0 |
| LRM Env 2 | N/A | N/A | 4.05 | 77.82 |
| LRM Env 3 | N/A | N/A | 32.19 | 101.26 |

Figure 4 reports the ET vs. GDT for our MicroTom in all three environments. The upper, middle, and lower sub-plots represent the greenhouse environments Env 1, Env 2, and Env 3, respectively. This time, the EM algorithm operated directly on the ET (and not on its reciprocal). In this scenario, measurement data were missing roughly 50% of the time. The start of the prediction was at $T = 10$. The prediction length $q = 1$. It can be seen that our predictor was able to follow the trend of the measurement curve. It predicted small outliers caused by changes in the environmental parameters with increasing accuracy the longer the observation time was. However, it had difficulties with anticipating large outliers. This was true for all environments. The Carmassi model in (2) was only quite accurate for Env 1, on which it was calibrated. The linear dependency of the LRM in (26) overemphasized the impact of the environmental parameters on the ET, in particular for Env 2.

Figure 5 illustrates the measured cumulative ET vs. the predicted for our predictor in (16). It can be seen that the predicted values were roughly Gaussian distributed around the 1:1 line, indicating freedom from bias. This was true for all three environments. The average prediction error of our DBN application was 29.42%.

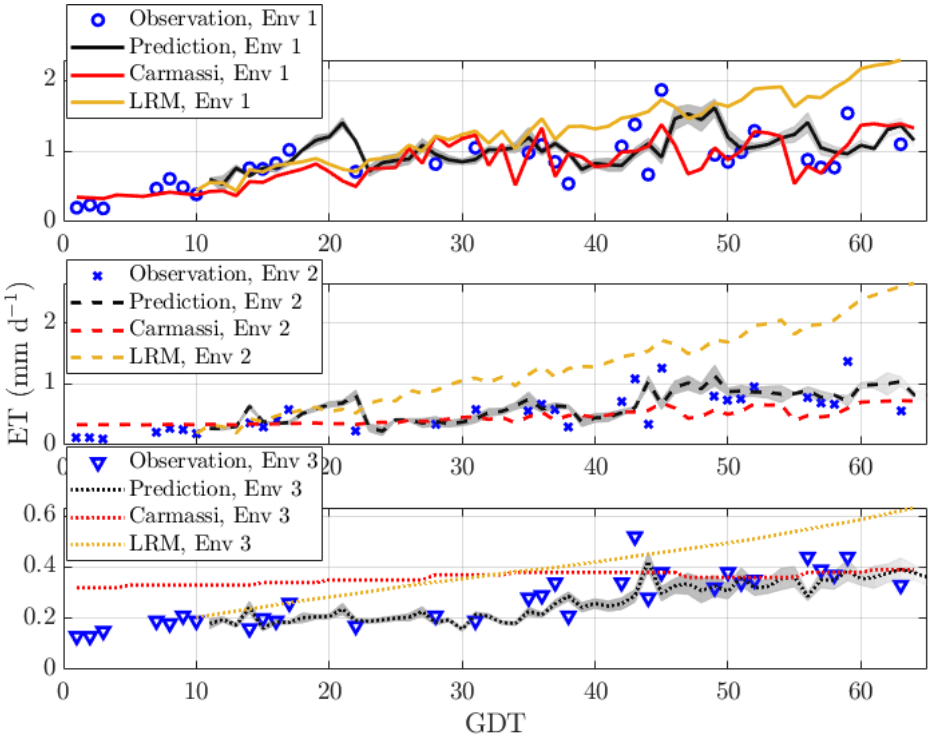


**Figure 4.** Prediction of the ET for MicroTom with sparse data in three environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination).

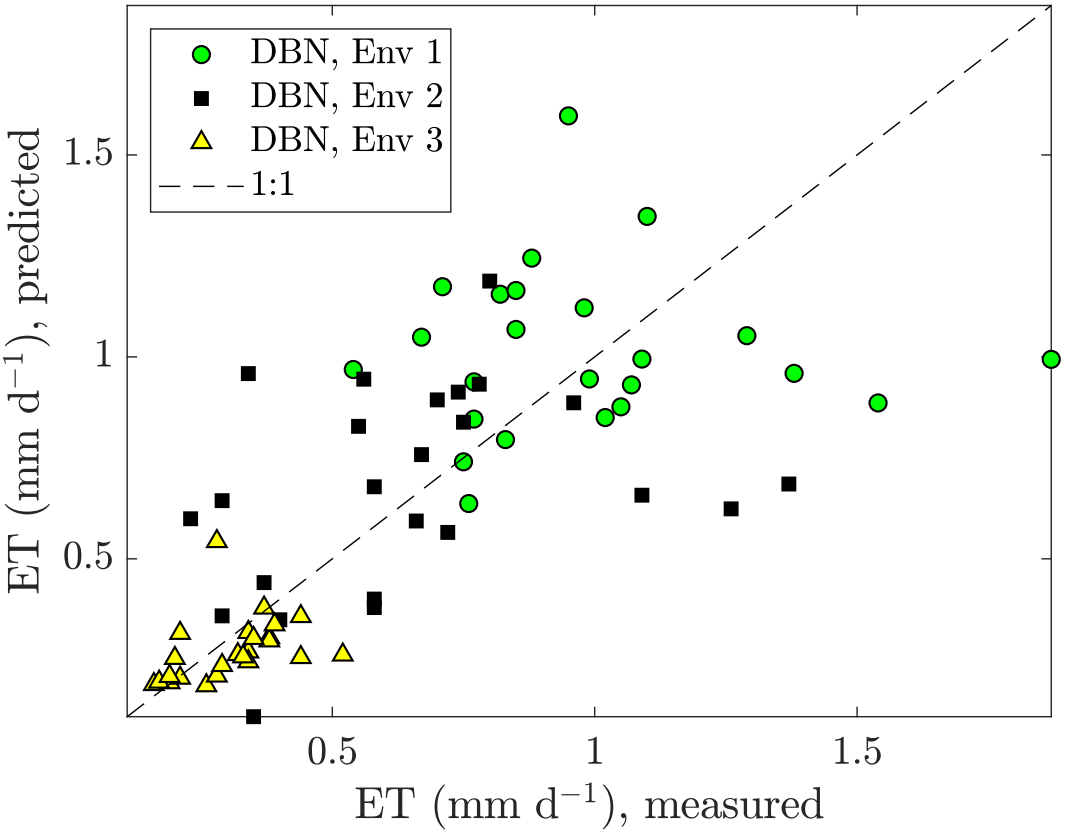


**Figure 5.** Measured vs. predicted measurement data obtained by our DBN application for the micro-tomatoes in three different environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination).

Not shown in the already overcrowded plot, the Carmassi model and the LRM had average prediction errors of 37.21% and 67.23%, respectively. Thus, it could be concluded that our DBN performs similar to the analytical Carmassi model. Error performance details are listed in Table 4.

**Table 4.** Estimation error of ET vs. GDT for the micro-tomatoes in three different environments: Env 1 (warm and bright), Env 2 (shaded and emergency heating), and Env 3 (indoor and artificial illumination).

| Error (%) | Env 1 | Env 2 | Env 3 |
|---|---|---|---|
| DBN | 30.13 | 37.54 | 20.57 |
| Carmassi | N/A | 37.85 | 36.57 |
| LRM | 49.56 | 112.21 | 39.92 |

## 4. Conclusions

We proposed a Bayesian machine learning approach for the inference of dynamical systems that modeled crop growth indicators with a sigmoid evolution over time. Such an evolution essentially corresponded to a two-phase dynamical process consisting of an exponential growth of the plant followed by a slow-down, leading to the achievement of a steady state. Both LAI and ET, which are among the most common growth-related indicators, exhibit such a dynamics for some types of plants.

The main advantage of our approach was flexibility. The model was inferred from the first items of the time series of indicators and environmental parameters, and it could be used to predict future items. This enabled real-time estimation of crop growth without the need to dedicate a whole cultivation cycle to a species-specific model calibration.

Compared to analytical models based on regression, which can fit calibration data in a very accurate way, our approach turned out to be more robust to changes in the environmental parameters. Again, this was a consequence of the fact that the model learned the influence of environmental parameters on the growth indicators from (the first items of) the same time series it would have to predict.

The proposed approach could also take advantage of available information about the usual trend of the growth indicators of interest. For instance, in the considered MicroTom tomatoes example, we used knowledge on the reciprocal LAI steady state (usually close to one) to initialize the matrices for the learning phase properly. This allowed us to obtain a more accurate inference of the dynamical model, and consequently more accurate predictions. Generalizing, this showed that the approach could also take available agronomic knowledge into account in order to improve the accuracy of predictions. For example, another piece of information that could be used, if available, was the usual time the steady state was expected to be reached (i.e., the expected length of the cultivation cycle). As future work, we plan to investigate how we could transform this information into an additional constraint for the learning phase and to measure the effect of such a constraint on the prediction capability of the inferred model.

**Author Contributions:** A.K.: conceptualization, methodology, software, writing, original draft. G.C.: data curation, validation, writing, review and editing. F.C.: data curation, validation. L.I.: Project administration, funding acquisition, writing, review and editing. P.M.: writing, review and editing S.C.: funding acquisition, writing, review and editing. All authors read and agreed to the published version of the manuscript.

**Funding:** This work is supported in part by the projects High-Tech House Garden (HT-HG) and COLTIV@MI funded by the Region of Tuscany under Bandi POR FESR 2014-2020, Bando 2.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| COLTIV@MI | COLTIVazione Automatizzata Miniaturizzata Innovativa |
| DBN | Dynamic Bayesian Network |
| EM | Expectation-Maximization |
| ET | Evapotranspiration |
| GDD | Growing Degree Days |
| GDT | Growing Days after Transplantation |
| ICT | Information and Communication Technology |
| IoT | Internet of Things |
| LAI | Leaf Area Index |
| MAR | Missing At Random |
| MCAR | Missing Completely At Random |
| RBF | Radial Basis Function |
| SVR | Support Vector Regression |

## Appendix A. Application of the EM Algorithm to Plant Growth Tracking

In this Appendix, we derive the EM algorithm for the state-space model in (6) and (7) when measurement data are missing at random.

From (10), the log-likelihood function of $\boldsymbol{\theta}$ for the complete data $\mathcal{X}$ is given by:

$$\begin{aligned}\ln p\left(\mathcal{X}|\boldsymbol{\theta}\right) \propto & -\ln|\boldsymbol{\Sigma}_1| - (\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) \\ & - T\ln|\boldsymbol{\Sigma}_w| - \sum_{t=1}^{T}(\boldsymbol{y}^{(\mathrm{obs})}t-\boldsymbol{C}\boldsymbol{x}_t)^T\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{y}^{(\mathrm{obs})}t-\boldsymbol{C}\boldsymbol{x}_t) \\ & - (T-1)\ln|\boldsymbol{\Sigma}_n| \\ & - \sum_{t=2}^{T}\left(\boldsymbol{x}_t-(\boldsymbol{I}-\boldsymbol{A}\,\boldsymbol{U}_{t-1})\,\boldsymbol{x}_{t-1}-\boldsymbol{B}\boldsymbol{u}_{t-1}\right)^T\boldsymbol{\Sigma}_n^{-1}\left(\boldsymbol{x}_t-(\boldsymbol{I}-\boldsymbol{A}\,\boldsymbol{U}_{t-1})\,\boldsymbol{x}_{t-1}-\boldsymbol{B}\boldsymbol{u}_{t-1}\right).\end{aligned} \tag{A1}$$

Substituting (A1) for the expected log-likelihood function function in (11), the result depends on the four conditional expectations:

$$\begin{aligned}\boldsymbol{x}_t^{[i]} &\triangleq \mathbb{E}\{\boldsymbol{x}_t|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\} \\ \boldsymbol{V}_t^{[i]} &\triangleq \mathrm{Cov}\{\boldsymbol{x}_t|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\} \\ \left(\boldsymbol{x}_t\boldsymbol{x}_t^T\right)^{[i]} &\triangleq \mathbb{E}\{\boldsymbol{x}_t\boldsymbol{x}_t^T|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\} = \boldsymbol{V}_t^{[i]}+\boldsymbol{x}_t^{[i]}\left(\boldsymbol{x}_t^{[i]}\right)^T \\ \left(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^T\right)^{[i]} &\triangleq \mathbb{E}\{\boldsymbol{x}_t\boldsymbol{x}_{t-1}^T|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\} = \boldsymbol{V}_t^{[i]}\boldsymbol{J}_{t-1}^T+\boldsymbol{x}_t^{[i]}\left(\boldsymbol{x}_{t-1}^{[i]}\right)^T,\end{aligned} \tag{A2}$$

with the short-cut:

$$\boldsymbol{J}_{t-1} \triangleq \mathrm{Cov}\{\boldsymbol{x}_{t-1},\boldsymbol{x}_t|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\}\mathrm{Cov}\{\boldsymbol{x}_t|\boldsymbol{y}^{(\mathrm{obs})},\boldsymbol{\theta}^{[i]}\}^{-1}. \tag{A3}$$

The forward-backward algorithm [44,45] can be used to efficiently compute the expectations in (A2). Following the approach in [19], the forward recursion of the forward-backward algorithm yields:

$$\begin{aligned}\boldsymbol{\mu}_t &\triangleq \boldsymbol{m}_{t-1}+\begin{cases}\boldsymbol{K}_t\left(\boldsymbol{y}_t-\boldsymbol{C}^{[i]}\boldsymbol{m}_{t-1}\right) & ; \quad \boldsymbol{y}_t\in\boldsymbol{y}^{(\mathrm{obs})} \\ \boldsymbol{0} & ; \quad \boldsymbol{y}_t\in\boldsymbol{y}^{(\mathrm{mis})}\end{cases} \\ \boldsymbol{V}_t &\triangleq \left(\boldsymbol{I}-\boldsymbol{K}_t\boldsymbol{C}^{[i]}\right)\left(\left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_{t-1}\right)\boldsymbol{V}_{t-1}\left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_{t-1}\right)^T+\boldsymbol{\Sigma}_n^{[i]}\right).\end{aligned} \tag{A4}$$

with the Kalman gain matrix $\boldsymbol{K}$:

$$\boldsymbol{K}_t \triangleq \boldsymbol{P}_{t-1}\left(\boldsymbol{C}^{[i]}\right)^T\left[\boldsymbol{C}^{[i]}\boldsymbol{P}_{t-1}\left(\boldsymbol{C}^{[i]}\right)^T+\boldsymbol{\Sigma}_{\boldsymbol{w}}^{[i]}\right]^{-1}, \tag{A5}$$

and the short-cuts:

$$\boldsymbol{m}_t \triangleq \left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_t\right)\boldsymbol{\mu}_t+\boldsymbol{B}^{[i]}\boldsymbol{u}_t \tag{A6}$$

and:

$$\boldsymbol{P}_t \triangleq \left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_t\right)\boldsymbol{V}_t\left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_t\right)^T+\boldsymbol{\Sigma}_{\boldsymbol{n}}^{[i]}. \tag{A7}$$

The backward recursion, in contrast, yields the state estimate of interest, namely a normal distribution with mean and covariance:

$$\begin{aligned}\boldsymbol{x}_t^{[i]} =&\boldsymbol{\mu}_t+\boldsymbol{J}_t\left(\boldsymbol{x}_{t+1}^{[i]}-\boldsymbol{m}_t\right),\\ \boldsymbol{V}_t^{[i]} =&\boldsymbol{V}_t+\boldsymbol{J}_t\left(\boldsymbol{V}_{t+1}^{[i]}-\boldsymbol{P}_t\right)\boldsymbol{J}_t^T,\end{aligned} \tag{A8}$$

respectively, where:

$$\boldsymbol{J}_t=\boldsymbol{V}_t\left(\boldsymbol{I}-\boldsymbol{A}^{[i]}\,\boldsymbol{U}_t\right)^T\boldsymbol{P}_t^{-1}. \tag{A9}$$

Finally, the boundary conditions are updated according to:

$$\begin{aligned}\boldsymbol{x}_T^{[i]}&=\boldsymbol{\mu}_T\\ \boldsymbol{V}_T^{[i]}&=\boldsymbol{V}_T.\end{aligned} \tag{A10}$$

The M-step of the EM algorithm in (11) learns the parameter vector $\boldsymbol{\theta}$ by computing the partial derivative of the expected log-likelihood function in (11), setting the result equal to zero, and solving with respect to the respective parameter, leaving:

$$\begin{aligned}\boldsymbol{B}^{[i+1]} =&\left[\sum_{t=2}^{T}\left(\boldsymbol{x}_t^{[i]}-\boldsymbol{x}_{t-1}^{[i]}\right)\boldsymbol{u}_{t-1}^T-\left(\left(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^T\right)^{[i]}-\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}\right)\boldsymbol{U}_{t-1}\right]\left[\sum_{t=2}^{T}\boldsymbol{U}_{t-1}\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}\boldsymbol{U}_{t-1}\right]^{-1}\\ &\times\left[\sum_{t=2}^{T}\boldsymbol{U}_{t-1}\boldsymbol{x}_{t-1}^{[i]}\boldsymbol{u}_{t-1}^T\right]\\ &\times\left[\sum_{t=2}^{T}\boldsymbol{u}_{t-1}\boldsymbol{u}_{t-1}^T-\sum_{t=2}^{T}\boldsymbol{u}_{t-1}\left(\boldsymbol{x}_{t-1}^T\right)^{[i]}\boldsymbol{U}_{t-1}\left(\sum_{t=2}^{T}\boldsymbol{U}_{t-1}\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}\boldsymbol{U}_{t-1}\right)^{-1}\sum_{t=2}^{T}\boldsymbol{U}_{t-1}\boldsymbol{x}_{t-1}^{[i]}\boldsymbol{u}_{t-1}^T\right]^{-1}\end{aligned} \tag{A11}$$

$$\begin{aligned}\boldsymbol{A}^{[i+1]} =&\left[\sum_{t=2}^{T}\left(\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}-\left(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^T\right)^{[i]}\right)\boldsymbol{U}_{t-1}+\boldsymbol{B}^{[i+1]}\boldsymbol{u}_{t-1}\left(\boldsymbol{x}_{t-1}^T\right)^{[i]}\boldsymbol{U}_{t-1}\right]\\ &\times\left[\sum_{t=2}^{T}\boldsymbol{U}_{t-1}\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}\boldsymbol{U}_{t-1}\right]^{-1}\end{aligned} \tag{A12}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{n}}^{[i+1]} =&\frac{1}{T-1}\sum_{t=2}^{T}\left(\boldsymbol{x}_t\boldsymbol{x}_t^T\right)^{[i]}-\left(\boldsymbol{I}-\boldsymbol{A}^{[i+1]}\boldsymbol{U}_{t-1}\right)\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_t^T\right)^{[i]}-\boldsymbol{B}^{[i+1]}\boldsymbol{u}_{t-1}\left(\boldsymbol{x}_t^T\right)^{[i]}\\ &+\left[\left(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^T\right)^{[i]}-\left(\boldsymbol{I}-\boldsymbol{A}^{[i+1]}\boldsymbol{U}_{t-1}\right)\left(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^T\right)^{[i]}-\boldsymbol{B}^{[i+1]}\boldsymbol{u}_{t-1}\left(\boldsymbol{x}_{t-1}^T\right)^{[i]}\right]\left(\boldsymbol{I}-\boldsymbol{U}_{t-1}\left(\boldsymbol{A}^T\right)^{[i+1]}\right)\\ &\left(-\boldsymbol{x}_t^{[i]}+\left(\boldsymbol{I}-\boldsymbol{A}^{i+1]}\boldsymbol{U}_{t-1}\right)\boldsymbol{x}_{t-1}^{[i]}+\boldsymbol{B}^{[i+1]}\boldsymbol{u}_{t-1}\right)\boldsymbol{u}_{t-1}^T\left(\boldsymbol{B}^T\right)^{[i+1]}\end{aligned} \tag{A13}$$

$$\boldsymbol{C}^{[i+1]} = \left( \sum_{t \in \text{obs}} \boldsymbol{y}_t^{(\text{obs})} \left( \boldsymbol{x}_t^{[i]} \right)^T \right) \left( \sum_{t \in \text{obs}}^{T} \left( \boldsymbol{x}_t \boldsymbol{x}_t^T \right)^{[i]} \right)^{-1}. \tag{A14}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}}^{[i+1]} = \frac{1}{T^{(\text{obs})}} \sum_{t=1}^{T} \boldsymbol{y}_t^{(\text{obs})} \left( \boldsymbol{y}_t^{(\text{obs})} \right)^T - \boldsymbol{y}_t^{(\text{obs})} \left( \boldsymbol{x}_t^T \right)^{[i]} \left( \boldsymbol{C}^T \right)^{[i+1]} - \boldsymbol{C}^{[i+1]} \left( \boldsymbol{x}_t^{[i]} \boldsymbol{y}_t^{(\text{obs})\,T} - \left( \boldsymbol{x}_t \boldsymbol{x}_t^T \right)^{[i]} \right) \left( \boldsymbol{C}^{[i+1]} \right)^T, \tag{A15}$$

$$\boldsymbol{\mu}_1^{[i+1]} = \boldsymbol{x}_1^{[i]}, \tag{A16}$$

$$\boldsymbol{\Sigma}_1^{[i+1]} = \left( \boldsymbol{x}_1 \boldsymbol{x}_1^T \right)^{[i]} - \boldsymbol{\mu}_1^{[i+1]} \left( \boldsymbol{\mu}_1^{[i+1]} \right)^T = \boldsymbol{V}_1^{[i]}. \tag{A17}$$

Note that all the parameters related to the measurements in (A13) and (A14) are based on the observed measurement data, ignoring the missing data. The above state and parameter estimates are the base of our DBN in Section 2.2.

## References


1. FAO. *The Future of Food and Agriculture: Trends and Challenges*; Food and Agriculture Organization of the United Nations (FAO): Rome, Italy, 2017.
2. Nicol, L.; Nicol, C. Adoption of precision agriculture to reduce inputs, enhance sustainabiltiy and increase food production: A study of southern Alberta, Canada. *WIT Trans. Ecol. Environ.* **2018**, *217*, 327–336. [CrossRef]
3. Gomez, C.; Chessa, S.; Fleury, A.; Roussos, G.; Preuveneers, D. Internet of Things for enabling smart environments: A technology-centric perspective. *J. Ambient Intell. Smart Environ.* **2019**, *11*, 23–43. [CrossRef]
4. Khanna, A.; Kaur, S. Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture. *Comput. Electron. Agric.* **2019**, *157*, 218–231. [CrossRef]
5. Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef]
6. Balducci, F.; Impedovo, D.; Pirlo, G. Machine learning applications on agricultural datasets for smart farm enhancement. *Machines* **2018**, *6*, 38. [CrossRef]
7. Rehman, T.U.; Mahmud, M.S.; Chang, Y.K.; Jin, J.; Shin, J. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Comput. Electron. Agric.* **2019**, *156*, 585–605. [CrossRef]
8. Burchi, G.; Chessa, S.; Gambineri, F.; Kocian, A.; Massa, D.; Milano, P.; Milazzo, P.; Rimediotti, L.; Ruggeri, A. Information Technology Controlled Greenhouse: A System Architecture. In Proceedings of the IoT Vertical and Topical Summit for Agriculture, Tuscany, Italy, 8–9 May 2018; IEEE: Tuscany, Italy, 2018. [CrossRef]
9. Mekonnen, Y.; Namuduri, S.; Burton, L.; Sarwat, A.; Bhansali, S. Review—Machine Learning Techniques in Wireless Sensor Network Based Precision Agriculture. *J. Electrochem. Soc.* **2020**, *167*, 037522. [CrossRef]
10. Fernandez, J.; Orsini, F.; Baeza, E.; Oztekin, G.; Muñoz, P.; Contreras, J.; Montero, J. Current trends in protected cultivation in Mediterranean climates. *Eur. J. Hortic. Sci.* **2018**, *83*, 294–305, [CrossRef]
11. Gruda, N.; Bisbis, M.; Tanny, J. Impacts of protected vegetable cultivation on climate change and adaptation strategies for cleaner production—A review. *J. Clean. Prod.* **2019**, *225*, 324–339. [CrossRef]
12. Müller, K.R.; Smola, A.J.; Rätsch, G.; Schölkopf, B.; Kohlmorgen, J.; Vapnik, V. *Predicting Time Series with Support Vector Machines*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 1997; Volume 1327, pp. 999–1004.
13. Mukerjee, S.; Osuna, E.; Girosi, F. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. In Proceedings of the IEEE Workshop on Neural Networks for Signal Processing VII, Amelia Island, FL, USA, 24–26 September 1997; pp. 511–520.
14. Jheng, T.Z.; Li, T.H.; Lee, C.P. Using Hybrid Support Vector Regression to Predict Agricultural Output. In Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC 2018), Hualien, Taiwan, 30 April–1 May 2018.
15. Zong, J.; Zhu, Q. Price forecasting for agricultural products based on BP and RBF Neural Network. In Proceedings of the 12th IEEE Int. Conf. Computer Science and Automation Engineering, Beijing, China, 22–24 June 2012; pp. 607–610.

16. Chapman, R.; Cook, S.; Donough, C.; Lim, Y.L.; Ho, P.V.V.; Lo, K.W.; Oberthür, T. Using Bayesian networks to predict future yield functions with data from commercial oil palm plantations: A proof of concept analysis. *Comput. Electron. Agric.* **2018**, *151*, 338–348. [CrossRef]
17. Carlson, G.A. A decision theoretic approach to crop disease prediction and control. *Am. J. Agric. Econ.* **1970**, *52*, 216–223. [CrossRef]
18. Bi, C.; Chen, G. Bayesian networks modeling for Crop Diseases. In Proceedings of the International Conference on Computer and Computing Technologies in Agriculture, Beijing, China, 27–30 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 312–320.
19. Kocian, A.; Massa, D.; Cannazzaro, S.; Incrocci, L.; Lonardo, S.D.; Milazzo, P.; Chessa, S. Dynamic Bayesian Network for Crop Growth Prediction in Greenhouses. *Comput. Electron. Agric.* **2020**. [CrossRef]
20. Rubin, D.B. Inference and Missing Data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
21. Yuan, Y.C. *Multiple Imputation for Missing Data: Concepts and New Development*; Technical Report; SAS Institute Inc.: Rockville, MD, USA, 2000.
22. Batista, G.; Monard, M. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* **2003**, *17*, 519–533. [CrossRef]
23. Takahashi, M.; Ito, T. Comparison of competing algorithms of multiple imputation: Analysis using large-scale economic data. *Res. Mem. Off. Stat.* **2014**, *71*, 39–82.
24. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser.* **1977**, *39*, 1–38.
25. Jiang, W.J.; Josse, J.; Lavielle, M.; TraumaBase, G. Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. *Comput. Stat. Data Anal.* **2020**, *145*. [CrossRef]
26. COLTIV@MI. Mini sErra Domestica Innovativa (Ital., Innovative Mini-Greenhouse). Available online: https://www.coltivami.com (accessed on 15 April 2020).
27. Scott, J.W.; Harbaugh, B.K. *Micro-Tom: A Miniature Dwarf Tomato*; Circular S-370; Agricultural Experiment Station, Institute of Food and Agricultural Sciences, University of Florida: Gainesville, FL, USA, 1989.
28. Carmassi, G.; Incrocci, L.; Maggini, R.; Malorgio, F.; Tognoni, F.; Pardossi, A. An aggregated model for water requirements of greenhouse tomato grown in closed rockwool culture with saline water. *Agric. Water Manag.* **2007**, *88*, 73–82. [CrossRef]
29. Scholberg, J.; McNeal, B.; Jones, J.; Boote, K.; Stanley, C.; Obreza, T. Growth and Canopy Characteristics of Field-Grown Tomato. *Agron. J.* **2000**, *92*, 152–159. [CrossRef]
30. Allen, R.G.; Pereira, L.S.; Raes, D.; Smith, M. *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage Paper 56*; Food and Agriculture Organization of the United Nations: Rome, Italy, 1998.
31. Stanghellini, C. Transpiration of Greenhouse Crops: An Aid to Climate Management. Ph.D. Thesis, Institute of Agricultural Engineering (IMAG), Wageningen, The Netherlands, 1987.
32. Katsoulas, N.; Stanghellini, C. Modelling Crop Transpiration in Greenhouses: Different Models for Different Applications. *Agronomy* **2019**, *9*, 392. [CrossRef]
33. Baille, M.; Baille, A.; Laury, J.C. A simplified model for predicting evapotranspiration rate of nine ornamental species vs. climate factors and leaf area. *Sci. Hortic.* **1994**, *59*, 217–232. [CrossRef]
34. Bailey, B.; Montero, J.; Biel, C.; Wilkinson, D.; Anton, A.; Jolliet, O. Transpiration of Ficus benjamina: Comparison of measurements with predictions of the Penman-Monteith model and a simplified version. *Agric. For. Meteorol.* **1993**, *65*, 229–243. [CrossRef]
35. Kittas, C.; Katsoulas, N.; Baille, A. Transpiration and canopy resistance of greenhouse soilless roses: Measurements and modelling. *Acta Hortic.* **1999**, *507*, 61–68. [CrossRef]
36. Montero, J.; Antón, A.; Muñoz, P.; Lorenzo, P. Transpiration from geranium grown under high temperatures and low humidities in greenhouses. *Agric. For. Meteorol.* **2001**, *107*, 323–332. [CrossRef]
37. Rouphael, Y.; Colla, G. Modelling the transpiration of a greenhouse zucchini crop grown under a Mediterranean climate using the Penman-Monteith equation and its simplified version. *Aust. J. Agric. Res.* **2004**, *55*, 931–937. [CrossRef]
38. Medrano, E.; Lorenzo, P.; Sánchez-Guerrero, M.; Montero, J. Evaluation and modelling of greenhouse cucumber-crop transpiration under high and low radiation conditions. *Sci. Hortic.* **2005**, *105*, 163–175. [CrossRef]

39. Carmassi, G.; Bacci, l.; Bronzini, M.; Incrocci, L.; Maggini, R.; Bellocchi, G.; Massa, D.; Pardossi, A. Modelling transpiration of greenhouse gerbera (Gerbera jamesonii H.Bolus) grown in substrate with saline water in a Mediterranean climate. *Sci. Hortic.* **2013**, *156*, 9–18. [CrossRef]
40. Bacci, L.; Battista, P.; Cardarelli, M.; Carmassi, G.; Rouphael, Y.; Incrocci, L.; Malorgio, F.; Pardossi, A.; Rapi, B.; Colla, G. Modelling Evapotranspiration of Container Crops for Irrigation Scheduling. In *Evapotranspiration—From Measurements to Agricultural and Environmental Applications*; IntechOpen: London, UK, 2011; Chapter 14, pp. 263–282. [CrossRef]
41. Wu, C.F.J. On the convergence properties of the EM algorithm. *Ann. Stat.* **1983**, *11*, 95–103. [CrossRef]
42. Hu, B.; Kocian, A.; Piton, R.; Hviid, A.; Fleury, B.H.; Rasmussen, L.K. Iterative joint channel estimation and interference cancellation using a SISO-SAGE algorithm for coded CDMA. In Proceedings of the 38th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 7–10 November 2004; pp. 622–626. [CrossRef]
43. Bacha, H.; Tekaya, M.; Drine, S.; Guasmi, F.; Touil, L.; Enneb, H.; Triki, T.; Cheour, F.; Ferchichi, A. Impact of salt stress on morpho-physiological and biochemical parameters of Solanum lycopersicum cv. Microtom leaves. *S. Afr. J. Bot.* **2017**, *108*, 364–369. [CrossRef]
44. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
45. Minka, T.P. *From Hidden Markov Models to Linear Dynamical Systems*; Technical Report TR-531; MIT Media Lab, MIT: Cambridge, MA, USA, 1999.