NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

MASTER THESIS

---

# Statistical Techniques for Hydroponic Greenhouse's Irrigation Management Driven By Functional-Structural Plant Models

---

*Author:*
Konstantinos FLORAKIS

*Supervisor:*
Samis TREVEZAS

Department of Mathematics

September 20, 2023

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

# *Abstract*

Department of Mathematics

Master In Statistical And Operational Research

**Statistical Techniques for Hydroponic Greenhouse's Irrigation Management Driven By Functional-Structural Plant Models**

by Konstantinos FLORAKIS

This master thesis presents an introduction to hydroponic greenhouse irrigation management, focusing on applying functional-structural plant models for water consumption estimation and prediction. Hydroponic greenhouse cultivation has become increasingly popular in recent years due to its numerous advantages over traditional farming methods, such as greater efficiency in resource use, higher crop yields, and the ability to grow plants year-round. However, precise and effective irrigation management is critical for achieving optimal crop growth and yield in hydroponic greenhouses.

To address this issue, functional-structural plant models, including the GreenLab model, are investigated to estimate and predict water consumption in hydroponic greenhouses with a particular focus on "Ekstasis" Tomato, where data are available. Through this research, innovative concepts emerge, leading to fresh approaches in the modeling. These models are fitted via maximum likelihood for parameter estimation. The results demonstrate that the above models provide a more accurate and precise approach to irrigation management, with biological interpretation, which significantly improves water use efficiency.

The study also investigates the effects of various environmental and crop factors on water consumption in hydroponic greenhouses, such as temperature, humidity, light intensity, and plant growth stage. By incorporating these factors into the models, a more comprehensive understanding of the irrigation requirements of hydroponic crops is obtained, enabling more precise and efficient irrigation management.

Overall, this thesis contributes to the hydroponic greenhouse irrigation management field by providing a systematic and data-driven methodology that can be applied in practical settings. The findings of this study may offer helpful insights into the sustainable cultivation of hydroponic crops, particularly in light of current global climate change concerns.

# Περίληψη

Στατιστικές Μέθοδοι στην Διαχείρηση Άρδευσης σε Υδροπονικά Θερμοκήπια, κατευθυνόμενες από λειτουργικά και δομικά μοντέλα

Η παρούσα διατριβή αποτελεί μια εισαγωγή στη διαχείριση της άρδευσης των υδροπονικών θερμοκηπίων, με έμφαση στην εφαρμογή λειτουργικών και δομικών μοντέλων για την εκτίμηση και πρόβλεψη της κατανάλωσης νερού. Τα μοντέλα που χρησιμοποιήθηκαν περιλαμβάνουν το GreenLab, τα οποία εφαρμόστηκαν μέσω της παραμετρικής μεθόδου μέγιστης πιθανοφάνειας. Κατα την διερεύνηση αυτών των μοντέλων, νέες μέθοδοι μοντελοποίησης του προβλήματος εμφανίστηκαν. Τα αποτελέσματα δείχνουν ότι τα μοντέλα αυτά παρέχουν μια πιο ακριβή, αποτελεσματική προσέγγιση, με βιολογική περιγραφή, για τη διαχείριση της άρδευσης στα υδροπονικά θερμοκήπια.

Στη έρευνα αναλύονται επίσης οι επιπτώσεις διαφόρων περιβαλλοντικών και καλλιεργειακών παραγόντων στην κατανάλωση νερού στα υδροπονικά θερμοκήπια, όπως η θερμοκρασία, η υγρασία, η ένταση του φωτός και το στάδιο ανάπτυξης των φυτών. Με την ενσωμάτωση αυτών των παραγόντων στα μοντέλα, επιτυγχάνεται μια πιο συνολική κατανόηση των απαιτήσεων άρδευσης των υδροπονικών καλλιεργειών, επιτρέποντας πιο ακριβή και αποτελεσματική διαχείριση της άρδευσης.

Εν κατακλείδι, η συγκεκριμένη έρευνα επιθυμούμε να συμβάλλει στον τομέα της διαχείρισης άρδευσης στα υδροπονικά θερμοκήπια παρέχοντας μια συστηματική μεθοδολογία που μπορεί να εφαρμοστεί σε πρακτικές καταστάσεις. Λαμβάνοντας υπόψη τις τρέχουσες προκλήσεις της παγκόσμιας ασφάλειας τροφίμων και της κλιματικής αλλαγής, θεωρούμε ότι τα ευρήματα αυτής της μελέτης μπορούν να προσφέρουν χρήσιμες προτάσεις για τη βιώσιμη καλλιέργεια των υδροπονικών καλλιεργειών.

# *Acknowledgements*

As this Master's program comes to an end, I would like to express my gratitude toward my parents, who planted in me the seed of curiosity; my teachers and professors, who irrigated this seed accordingly; and my family and friends, who accept me for who I am.

# Contents

# Chapter 0

# Motivation and a bit of backstory

## 1 Greenhouse Irrigation for Sustainable Agriculture

Agriculture has been a key aspect of human civilization since the Neolithic revolution, serving as a major driving force in the food supply chain (Flannery, 1973). With advancements in science and technology, new tools and methods are being developed to increase the efficiency and precision of agricultural practices. One such tool is the greenhouse, which enables the growth of plants in regulated climatic conditions, even out of season (Oxford-University-Press, 1989). While various operations are crucial for greenhouse management, including ventilation, heating, cooling, and lighting, irrigation management (IM) plays a vital role in ensuring efficient plant growth and business operations.

IM in greenhouses involves monitoring and regulating water consumption, ensuring the appropriate amount of water is applied to plants while minimizing excess usage. Effective IM can prevent economic losses due to overwatering or water deficits and improve the overall efficiency of water usage (Ünlü et al., 2011). Therefore, it is essential to consider the water consumption of plants and the efficiency of water usage in IM practices.

In conclusion, implementing effective IM practices is critical for achieving efficient plant growth and sustainable business practices for greenhouse operations. The focus on efficient water usage will ensure successful plant growth and promote environmental sustainability in agriculture.

## 2 Backstory

In 2020, the world was confronted with a myriad of unprecedented challenges, most notably the global pandemic. As for me, I was in the process of completing the final courses of my Bachelor's degree when September arrived. I was confident in my ability to successfully finish these courses, but I found myself uncertain about my future path and aspirations.

In the midst of this uncertainty, I reached out to my uncle who lives in Serres and runs a greenhouse company. With the pandemic-imposed restrictions looming, I wanted to contribute to his business

beyond the traditional laborer roles. I was determined to find a problem that could put my knowledge and skills to good use. Serendipitously, the greenhouse, a family connection, and my desire to support my uncle created the perfect environment to uncover such a challenge.

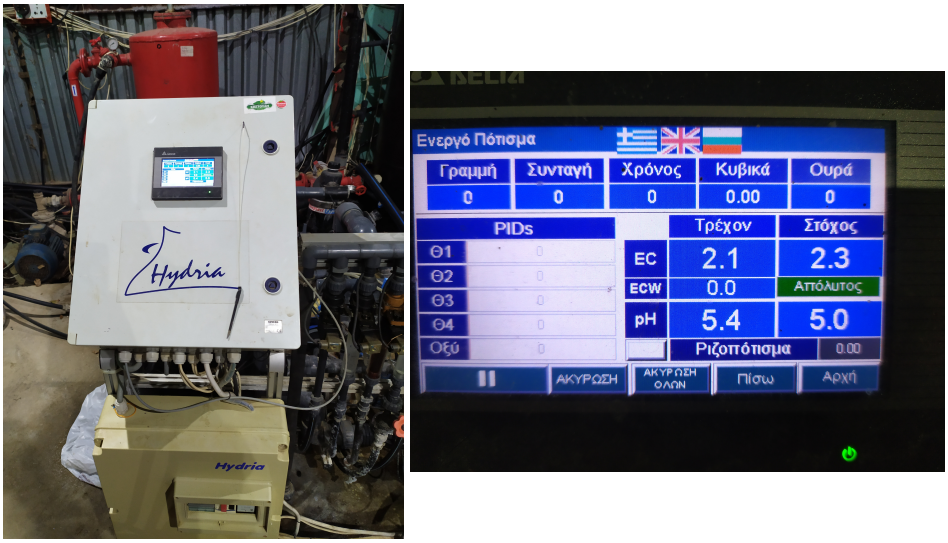


FIGURE 1: Irrigation Machine and menu

As fate would have it, my intuition was correct. I moved to Serres on the 8th of October 2020 and started working at my uncle's greenhouse company. In the early days, I observed my uncle and boss struggling to manage the irrigation system. The exhausting task of constantly monitoring the plants and adjusting the irrigation according to fluctuating weather conditions seemed both exasperating and time-consuming. Although I had no prior expertise in statistical modeling, I recognized this as the problem I sought to solve and eagerly began gathering data.

Fortunately, a meteorological station was already set up near the greenhouse, and my suggestion to install gauges inside the greenhouse to track temperature and humidity was well-received. However, the initial datasets we collected turned out to be flawed. Our initial method of recording water usage at 2:00 p.m. resulted in a one-day time interval, which merged readings from different days, like sunny and cloudy days, which could affect plant water consumption differently. Despite a month of hard work spent on these flawed datasets, I learned valuable lessons from the experience.

FIGURE 2: A station to measure Water Consumption

As time went by, I discovered a seminar on Monte-Carlo techniques led by Professor Trevezas. During the first session, the professor mentioned his previous work on plant modeling, which immediately captured my attention. I eagerly reached out to him through email, and our correspondence began. He generously provided me with guidance and suggestions for my early modeling attempts.

In January 2021, we implemented an initial irrigation regimen for tomato and cucumber crops based on a linear regression model. Over the following months, we set up a pipeline to incorporate data from the meteorological station into our models. After dividing the day into heuristic segments, we applied the generated regimen to the plants, resulting in production levels that met our expectations. Moreover, the supervising agronomist confirmed that key solution statistics like EC[1] and pH[2] remained stable. However, I must emphasize that these findings are purely anecdotal. This thesis aims to elevate this work to rigorous scientific standards.

[1] electrical conductivity; a non-specific measurement of the concentration of both positively and negatively charged ions within a sample, usually expressed in milli-Siemens per centimeter (mS/cm) or microSiemens per cm (μS/cm)

[2] the measure of hydrogen ion concentration in a sample used to determine the acidity or alkalinity of a product

# Chapter 1

# Data Driven Models (DDMs)

## 1 Linear Regression

The first model approach for the problem was Linear Regression (LR), mainly for simplicity. In the next paragraph, we recall LR, and then specialize it to our case. For more information, the reader can refer to (Faraway, 2002) or other standard books in LR.

### 1.1 The Linear hypothesis

Let $Y$ be the *dependent variable* (output) and $X_1, X_2, \ldots, X_p$ be the *independent variables* (input, predictors) for a given problem. The general problem of regression consists of finding a relationship of the form

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon,$$

by estimating from the output an appropriate function $f$ linking the output to the predictors while adding an error term $\varepsilon$ to accommodate for the discrepancies.

The linear hypothesis simplifies the problem to the following parameter estimation problem:

$$Y = b_0 + b_1 X_1 + \ldots + b_p X_p + \varepsilon,$$

where $b_0, b_1, \ldots, b_p$ are unknown coefficients and $b_0$ is typically called *intercept*.

Even though parameters enter linearly, the predictors do not have to be linear. For example:

$$Y = b_0 + b_1 \log X_1 + b_2 X_2^2 + \ldots + b_p X_p,$$

is still a linear model, but with respect to $\{1, \log X_1, X_2^2, \ldots, X_p\}$ and not the original predictor variables.

Linear models seem rather restrictive but because the predictors can be transformed and combined in any way, they are actually very flexible. Truly non-linear models often arise from a theory about relationships between variables rather than an empirical investigation.

## 1.2 Matrix Representation

The linear equations

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i, \qquad i \in \{1, \ldots, n\}$$

can be formulated as:

$$Y = Xb + \varepsilon, \tag{1.1}$$

where

$$Y = (y_1, \ldots, y_n)^T, \qquad b = (b_0, \ldots, b_p)^T, \qquad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$$

and the so-called design matrix is given by

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

where a column of ones was introduced to represent the intercept column. As we discussed before, the goal is to estimate $b$, in a way that minimizes $\varepsilon$ in an appropriate sense. Working on $\mathbb{R}^n$ euclidean distance will be used as the minimization criterion.

## 1.3 Least squares estimation

We might define the best estimate of $b$ as that which minimizes the sum of the squared errors, $\varepsilon^T\varepsilon$. That is to say that the least squares estimate of $b$, called $\hat{b}$ minimizes:

$$\begin{aligned} \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T\varepsilon &= (Y - Xb)^T(Y - Xb) \\ &= Y^TY - 2b^TX^TY + b^TX^TXb. \end{aligned}$$

Differentiating with respect to $b$ and setting to zero, we get that $\hat{b}$ satisfies:

$$-2X^TX\hat{b} + 2X^TY = 0$$
$$X^TX\hat{b} - X^TY = 0.$$

These are called the *normal equations*. Under the assumption that $X^TX$ is invertible:

$$\hat{b} = (X^TX)^{-1}X^TY.$$

After estimating $b$ with $\hat{b}$ we say that we have a model *fit* with the fitted values $\hat{Y}$ (predicted values of $Y$ based on the same training set):

$$\hat{Y} = X\hat{b} = X(X^TX)^{-1}X^TY = HY,$$

where $H = X(X^TX)^{-1}X^T$ is called the *hat matrix* and has some nice properties:

- H is symmetric ($H^T = H$)
- H is idempotent ($H^2 = H$)
- Predicted values: $\hat{Y} = HY = X\hat{b}$
- Residuals: $\hat{\varepsilon} = Y - X\hat{b} = Y - \hat{Y} = (I - H)Y$
- Residual sum of squares: $\hat{\varepsilon}^T\hat{\varepsilon} = Y^T(I - H)(I - H)Y = Y^T(I - H)Y$.

### 1.4 Gauss Markov Theorem

We found a way to estimate $b$ with $\hat{b}$. Why is $\hat{b}$ a good estimate? Actually there are two main reasons:

- If the errors are independent and identically normally distributed, it is the maximum likelihood estimator. Loosely put, the maximum likelihood estimate is the value of $b$ that maximizes the probability of the data that was observed.
- The Gauss-Markov theorem states that it is the best linear unbiased estimate.

**Theorem 1.1.** *Suppose $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{V}ar[\varepsilon] = \sigma^2 I$. Suppose also that the structural part of the model, $\mathbb{E}[Y] = Xb$ is correct. Let $\psi = c^Tb$, where there exists a linear combination $q^TY$ such that $\mathbb{E}[q^TY] = c^Tb$ (such $\psi$ is called* estimable*). Then, in the class of all unbiased linear estimates of $\psi$, $\hat{\psi} = c^T\hat{b}$ has the minimum variance and is unique.*

*Proof.* Let $q^T Y$ be a linear unbiased estimator of $\psi = c^T b$. We will show that $\mathbb{V}ar[q^T Y] \geq \mathbb{V}ar[c^T \hat{b}]$. Note here that $\hat{b} = (X^T X)^{-1} X^T Y$, in the case where $X^T X$ is invertible, which means that $\hat{b}$ is also a linear combination of $Y$.
Since $q^T Y$ is unbiased estimator of $c^T b$ we have for all $b$:

$$\begin{aligned} \mathbb{E}q^T Y &= q^T \mathbb{E}Y \\ \text{or} \quad q^T X b &= c^T b. \end{aligned} \tag{1.2}$$

Because this holds for all $b$, we derive $q^T X = c^T$. Another useful observation we have to note is that:

$$\begin{aligned} \mathbb{V}ar(c^T \hat{b}) &= c^T \mathbb{V}ar(\hat{b}) c \\ &= c^T (X^T X)^{-1} X^T \mathbb{V}ar(Y) \left( (X^T X)^{-1} X^T \right)^T c \\ &= \sigma^2 c^T (X^T X)^{-1} c, \end{aligned} \tag{1.3}$$

where we used the fact that because $X^T X$ is symmetric, its inverse (if exists) is also symmetric.

The last result we will need is referring to the *hat* matrix H, and it states that for any vector $v \in \mathbb{R}^{n \times 1}$:

$$v^T v \geq v^T H v. \tag{1.4}$$

To prove that, observe:

$$(I - H^T)H = (I - H)H = H - H^2 = 0, \quad \text{the zero vector.}$$

Now, the left hand side of 1.4 can be expressed as $||v||^2$, while the right hand side can be written as:

$$\begin{aligned} v^T H v &= v^T H^T H v, \quad \text{H idempotent and symmetric} \\ &= ||Hv||^2 \end{aligned}$$

By the Pythagorean theorem, we have:

$$||v||^2 = ||Hv||^2 + ||(I - H)v||^2, \tag{1.5}$$

which because $||(I-H)v||^2 \geq 0$ proves the claim.
Combining the previous observations, for the $\mathbb{V}ar(q^TY)$ we have:

$$\begin{aligned}
\mathbb{V}ar(q^TY) &= q^T\mathbb{V}ar(Y)q \\
&= \sigma^2 q^T q \\
&\geq \sigma^2 q^T H q \\
&= \sigma^2 q^T X(X^TX)^{-1}X^T q \\
&= \sigma^2 c^T(X^TX)^{-1}c \\
&= \mathbb{V}ar(c^T\hat{b}),
\end{aligned}$$

and the proof is complete. □

### 1.5 Mean and Variance of $\hat{b}$

Now $\hat{b} = (X^TX)^{-1}X^TY$ so its mean vector and variance(-covariance) matrix is given by

- Mean $\mathbb{E}[\hat{b}] = (X^TX)^{-1}X^TXb = b$ (unbiased)
- Variance $\mathbb{V}ar[\hat{b}] = \sigma^2(X^TX)^{-1}$.

### 1.6 Estimating $\sigma^2$

Recall that the residual sum of squares was $\hat{\varepsilon}^T\hat{\varepsilon} = Y^T(I-H)Y$. Now after some calculation, one can show that $\mathbb{E}[\hat{\varepsilon}^T\hat{\varepsilon}] = \sigma^2(n-p)$ which shows that

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p}$$

is an unbiased estimate of $\sigma^2$. As $b \in \mathbb{R}^{p\times 1}$ and $y \in \mathbb{R}^{n\times 1}$, the model has a systematic structure (after estimating $b$) over $p$ dimensions, with the $n-p$ dimensions that remain to be responsible for the random variation, there are, as a result, the degrees of freedom of the model.

## 1.7 Goodness of fit

### 1.7.1 $R^2$

How well does the model fit the data? The $\hat{Y} = (\hat{y}_i)_{i=1,\dots,n}$ is the vector of fitted values. One measure is $R^2$, the so-called coefficient of determination or percentage of variance explained

$$R^2 = 1 - \frac{RSS}{TSS},$$

where $RSS = \varepsilon^T \varepsilon$ is the residual sum of squares, and $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares. Usually $R^2$ is in $(0, 1)$. Values closer to 1 indicate better fits. We can also achieve negative values with really bad fits. For simple linear regression $R^2 = r^2$ where r is the correlation between x and y. One may face two different scenarios when attempting to predict the dependent variable y. In the absence of the independent variable x, the most reasonable prediction would be the mean of y, denoted as $\bar{y}$. However, this prediction may exhibit considerable variability due to the lack of information about x. Conversely, when x is known, the regression fit can be used to make more precise predictions, assuming there is some relationship between x and y. The coefficient of determination, $R^2$, quantifies the proportion of the variance in y that is predictable from x, with values closer to one indicating better predictions.

It is essential to note that the conventional definition of $R^2$ assumes the presence of an intercept in the model. In the absence of an intercept, alternative definitions of $R^2$ may be employed, but these should not be directly compared with the $R^2$ values obtained from models with an intercept. Caution is advised when interpreting high $R^2$ values derived from models without an intercept.

The acceptability of a given $R^2$ value is contingent upon the specific field of application. In the biological and social sciences, weaker correlations and increased noise are common, resulting in generally lower $R^2$ values. An $R^2$ value of 0.6 may be deemed satisfactory in these fields. On the other hand, in physics and engineering, controlled experimental conditions often yield higher $R^2$ values, with 0.6 considered relatively low. Ultimately, understanding the particular domain is necessary to assess the adequacy of an obtained $R^2$ value accurately.

### 1.7.2 $R^2$ adjusted

Even though $R^2$ has a very intuitive representation, it has a drawback. The $R^2$ statistic increases with respect to the number of parameters, which means that it can be misleading for a big number $p$ of

parameters. To avoid such pit holes, $R^2$-*adjusted* has been proposed:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \cdot \frac{n - 1}{n - p - 1},$$

where an adjustment has been made to penalize extra parameters.

### 1.7.3 Residual standard error

An alternative measure of model fit is the residual standard error, denoted as $\hat{\sigma}$. This statistic is directly related to the standard errors of the estimates of the model parameters and predictions. One of the advantages of using $\hat{\sigma}$ is that it is expressed in the same units as the response variable, facilitating interpretation within the context of the specific dataset. However, this property may also be considered a disadvantage, as it requires a deeper understanding of the practical significance of the measure, unlike the unitless $R^2$, which is more straightforward to comprehend.

## 2 Model Selection

In most practical settings, various regression models will be viable for our dataset. It's even plausible that we've broadened the scope of potential models by incorporating fresh variables derived from the ones initially accessible, achieved through alterations, introduction of interactions, or inclusion of polynomial expressions. This section delves into the difficulties when handpicking the optimal subset of predictors.

Additional information could seemingly be advantageous, leading one to ponder why we wouldn't just incorporate all accessible variables into the model. Nevertheless, opting for a more compact model may be preferable for several reasons. The principle of Occam's Razor asserts that the simplest one reigns supreme among various credible interpretations of a phenomenon. When applied to regression analysis, this conveys that the smallest model sufficiently accommodating the data is optimal.

Occam's Razor stands strong as a persuasive heuristic, yet our core focus within the realm of regression modeling must remain tireless. The potential for attaining heightened predictive capabilities lies in the realm of more comprehensive models. Thus, even though the concept of slimmer models may exude appeal, we remain resolute in our commitment to safeguarding predictive ability.

In the pursuit of comprehending the explanatory influence of predictors, exercising caution with regard to automated variable selection methods becomes necessary. These scenarios underscore the prominence of a handful of noteworthy predictors while relegating the remaining to auxiliary roles

demanding meticulous oversight. To place dependable responsibility upon an automated procedure for such a task would not be advised.

In the evaluation of prospective models, we could employ hypothesis testing techniques to facilitate selection or opt for criterion-based methods to gauge relative fit, guiding our decision-making process. These choices are both explored in the present section.

## 2.1 Testing-Based Procedures

Among variable selection techniques, **Backward Elimination** is the most straightforward and can be implemented without specialized software. Even in scenarios of intricate hierarchies, the manual execution of backward elimination remains feasible, factoring in the eligibility of variables for removal.

The procedure commences with all predictors within the model, subsequently discarding the predictor with the highest p-value surpassing the threshold $\alpha_{crit}$. The model is then reconfigured, followed by the exclusion of the least significant remaining predictor, under the condition that its p-value surpasses $\alpha_{crit}$. Each iteration excludes progressively more "nonsignificant" predictors until the selection process culminates.

The threshold $\alpha_{crit}$, at times referred to as "p-to-remove," need not be constrained to 5%. If the aim is predictive performance, a cutoff of 15 to 20% might yield optimal results, although methods specifically tailored for superior predictive accuracy are recommended. **Forward Selection**, in essence, operates in the reverse direction of the backward method. It commences with an absence of variables in the model. Then, their p-values are evaluated upon inclusion for each predictor not presently within the model. The predictor exhibiting the lowest p-value below the threshold $\alpha_{crit}$ is incorporated. This process persists until no additional predictors can be introduced. **Stepwise Regression** amalgamates backward elimination and forward selection elements. This approach is apt for circumstances where the inclusion or exclusion of variables is modified during the early phases, with the potential for alterations later on. A variable can be appended or removed at each stage, and diverse methodologies exist to execute these actions precisely.

Testing-oriented methodologies exhibit computational efficiency and straightforwardness, yet they carry with them significant limitations:

- The gradual process of adding or discarding variables, one by one, can result in overlooking the "optimal" model.
- Interpreting p-values too rigorously is cautioned due to extensive multiple testing, casting skepticism on their validity. The act of removing less impactful predictors frequently inflates the

significance of the remaining ones. This phenomenon fosters an exaggerated estimation of the retained predictors' significance.

- These procedures lack a direct link to ultimate objectives like prediction or explanation, potentially falling short in addressing the core problem. It's essential to bear in mind that, within any variable selection approach, selecting a model is inherently tied to the fundamental purpose of investigation. This process tends to magnify the statistical significance of variables retained within the model. Variables that are excluded may still exhibit correlations with the response variable. Concluding that these variables hold no relation to the response would be an oversimplification; their omission merely signifies they don't contribute additional explanatory power beyond the variables already encompassed in the model.

Stepwise variable selection often leans towards models that are more compact than ideal for predictive purposes. Imagine a basic regression featuring just one predictor variable. If the statistical significance of the slope for this predictor hovers near the threshold, there might not be substantial evidence to establish its association with $y$ firmly. However, it could still be judicious to leverage it for predictive intents.

## 2.2 Criterion-Based Procedures

When we clearly understand the intended purpose behind a model, we can establish a measure to assess how effectively a specific model aligns with that objective. From the range of possible models, we can then select the one that best optimizes this criterion.

Opting for a model $g$, parameterized by parameters $\theta$, that closely approximates the true model $f$ seems intuitive. The disparity between $g$ and $f$ could be quantified by calculating the distance using a function $L(f, g_\theta)$. Such a function is called a *loss* function. Many choices exist in the literature, we just note some of them here ($\mathcal{L}_g$ denotes the likelihood with respect to model $g$):

- AIC (Akaike Information Criterion):

$$\text{AIC}(f, g) = -2\,\log(\mathcal{L}_g) + 2\,k$$

- BIC (Bayesian Information Criterion):

$$\text{BIC}(f, g) = -2\,\log(\mathcal{L}_g) + k\,\log(n)$$

- Kullback-Leibler Divergence:

$$\mathrm{KL}(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

- Mean Squared Error (MSE):

$$\mathrm{MSE}(f, g) = \int (f(x) - g(x))^2 \, dx$$

Unfortunately, some of them (e.g., MSE) are impractical for direct implementation because we do not know $f$. To select among a set of models, we choose the model which minimizes the contextual loss function. All the stepwise methods mentioned in the previous section can be adapted in such a way that each step (addition or deletion) is chosen according to a loss function.

## Chapter 2

# Agronomic Variables

'Keep your face to the sun and you will never see the shadows.'
— Helen Keller.

The intricacies of agricultural ecosystems are often a confluence of multiple factors that determine the success and sustainability of crop production, efficiency of water consumption, and balanced growth. Among these factors, agronomic variables play a crucial role in shaping the productivity, quality, and resilience of agricultural systems. This chapter delves into the significance of several key agronomic variables, namely temperature, humidity, solar radiation, vapor pressure deficit, evapotranspiration, and others, in the context of modern agriculture.

A profound understanding of these variables is instrumental in designing effective agronomic management strategies that adapt to the dynamic and complex nature of agroecosystems (Hatfield and Prueger, 2015). The assessment of these factors provides essential insights into crop growth, development, and yield response under varying environmental conditions (Levitt (1980); Taiz et al. (2015)). Moreover, the comprehensive analysis of these agronomic variables helps in optimizing resource use, minimizing environmental impacts, and enhancing overall agricultural sustainability (Fageria, 2012).

In this chapter, we will explore each agronomic variable individually, discussing their effects on crop growth and development, as well as their implications for agricultural management practices.

## 1 Solar irradiance

Solar radiation, also referred to as solar energy or sunlight, is the electromagnetic energy emitted by the sun that reaches Earth's surface (Gueymard, 2004). The electromagnetic spectrum encompasses a broad range of wavelengths and frequencies. Ultraviolet (UV) radiation typically spans wavelengths from about 10 nm to 400 nm, with corresponding frequencies in the range of approximately 30 PHz (petahertz) to 750 THz (terahertz). Visible light lies between the ultraviolet and infrared regions, with

wavelengths ranging from around 400 nm (violet) to 700 nm (red) (Figure 2.2) and frequencies from approximately 430 THz (red) to 790 THz (violet). Infrared (IR) radiation extends from about 700 nm to 1 mm wavelengths, with corresponding frequencies ranging from roughly 430 THz down to 300 GHz (gigahertz). This radiant energy originates from nuclear fusion reactions taking place in the Sun's core, where hydrogen atoms merge to create helium, liberating an immense quantity of energy as electromagnetic radiation (Carroll and Ostlie, 2006).

Solar radiation plays a crucial role in driving various physical, chemical, and biological processes on Earth, including photosynthesis in plants, oceanic and atmospheric circulation, and the generation of weather patterns. It is also the primary source of renewable energy, with technologies such as photovoltaic cells and solar thermal collectors designed to harness solar energy for electricity generation and heating purposes respectively (Kalogirou (2004)).

**Definition 2.1.** *The per unit time quantity of solar radiation incident upon a unit area on Earth's surface is termed solar irradiance, which is typically measured in Watts per square meter ($W/m^2$). (Kopp and Lean, 2011)*

Different light wavelengths have different energy levels and, consequently, different irradiance values ($W/m^2$). The energy of a photon is proportional to its frequency and inversely proportional to its wavelength. As a result, shorter wavelengths (like ultraviolet) carry more energy per photon than longer wavelengths (like infrared).

What we mean in Definition 2.1 by "per unit time quantity of solar radiation" is actually referring to the power of the solar radiation. Solar irradiance involves the sum of the energy carried by all wavelengths of light (UV, visible, and IR) in the solar spectrum that reaches a given area on Earth's surface, usually expressed in Watts per square meter ($W/m^2$).

It is important to note that solar irradiance varies depending on factors such as the time of day, latitude, season, and atmospheric conditions (e.g., cloud cover, air pollution) (Kopp and Lean, 2011). This noise can also be observed in Figure 2.1. Observe the periodic variations in the data, which can be attributed to the Earth's movement and are responsible for the seasonal changes we experience throughout the year. As a result, the amount of solar radiation reaching any particular location on Earth's surface can change throughout the day and year.

Understanding solar radiation is essential for various scientific disciplines, such as climatology, meteorology, and renewable energy research, as it profoundly impacts the Earth's energy balance, climate systems, and the development of sustainable energy solutions.

Solar radiation is introduced as a key driving factor influencing evapotranspiration, a variable that will be elaborated upon in subsequent sections, and it will be represented by solar irradiance measured in $(W/m^2)$.

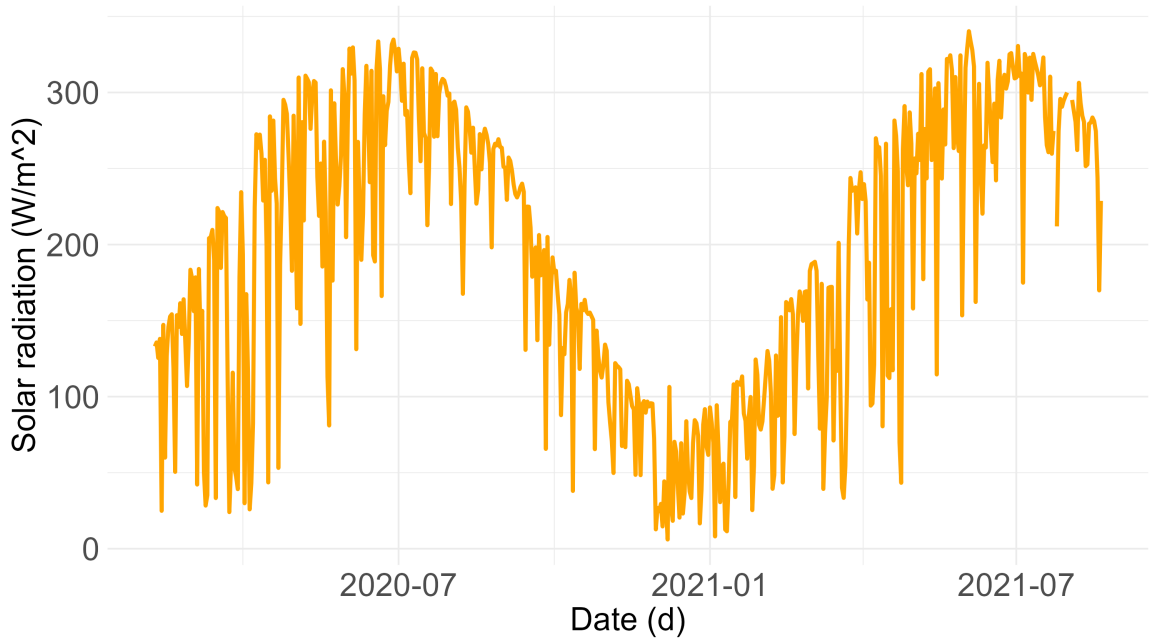


FIGURE 2.1: Solar radiation measurements from 2020 and 2021 are displayed, as recorded by a Davis Vantage Pro 2 (Plus) meteorological station.

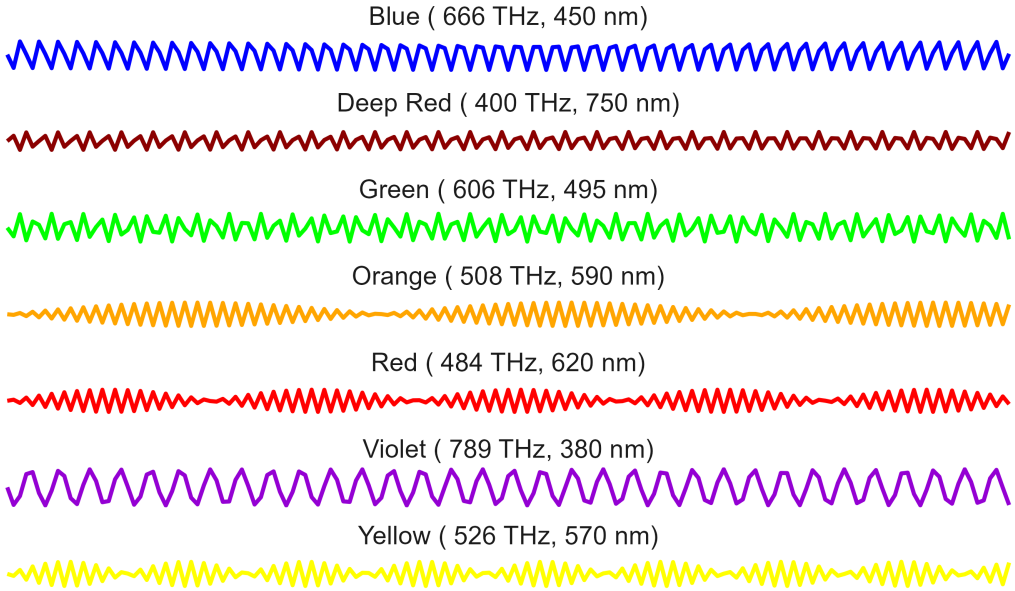


FIGURE 2.2: Illustration of the sinusoidal nature of light waves across the visible spectrum. The facet labels provide the corresponding frequency and wavelength ranges for each color. Note that this graph serves as a conceptual representation and does not directly depict the actual differences in wavelengths and frequencies between colors.

## 2 Temperature

**Definition 2.2.** *In scientific terms, temperature is a physical property that quantifies the degree of hotness or coldness of a substance or environment (Boltzmann, 2012). It serves as an essential parameter in thermodynamics and is commonly expressed in units such as Celsius (°C), Fahrenheit (°F), or Kelvin (K).*

The concept of temperature arises from the fundamental principle that, at the microscopic level, the kinetic energy of particles in a substance is directly related to their thermal agitation, with higher temperatures corresponding to greater particle motion (Fermi, 1956).

Temperature plays a critical role in various natural and artificial systems, including biological processes and industrial applications. In the context of plant growth and development, temperature significantly influences the physiological and biochemical activities of plants, with each species exhibiting a specific temperature range characterized by minimum, maximum, and optimum values (Boote et al., 2013). These temperature thresholds directly affect the plants' metabolic rates, nutrient uptake, photosynthesis, and overall growth performance (Porter and Gawith, 1999).

Furthermore, temperature is a vital factor in agricultural production, as it is one of the few environmental variables that can be controlled, albeit within certain limits. By manipulating temperature conditions, crop yields and quality can be optimized, making temperature regulation a crucial aspect of modern agricultural practices (Lobell et al., 2011).

In our context Temperature is also a driving factor of evapotranspiration, as is Solar radiation, and will be measured in Celsius ($^\circ C$).

## 3 Thermal time

Temperature is widely considered a critical environmental factor in agriculture, significantly influencing plant growth and development (Boote et al., 2013). Plant maturation relies on adequate heat, and extreme temperature conditions can induce stress and impede growth rates. In agricultural research, this strong temperature dependence is often modeled using Growing Degree Days (GDD), which quantifies the heat received by vegetation in a single day as a time unit for estimating crop growth (Snyder et al., 1999).

The underlying principle of GDD is that a crop's growth rate optimizes at a specific temperature ($T_o$) and decreases to zero below a base temperature ($T_b$) or above a ceiling temperature ($T_c$). Consequently, the average daily temperature can be compared against these extreme temperatures to serve as a growth indicator. The summation of GDD over a plant's lifespan yields the Accumulated Growing Degree Days (AGDD), a statistic representing the plant's age.

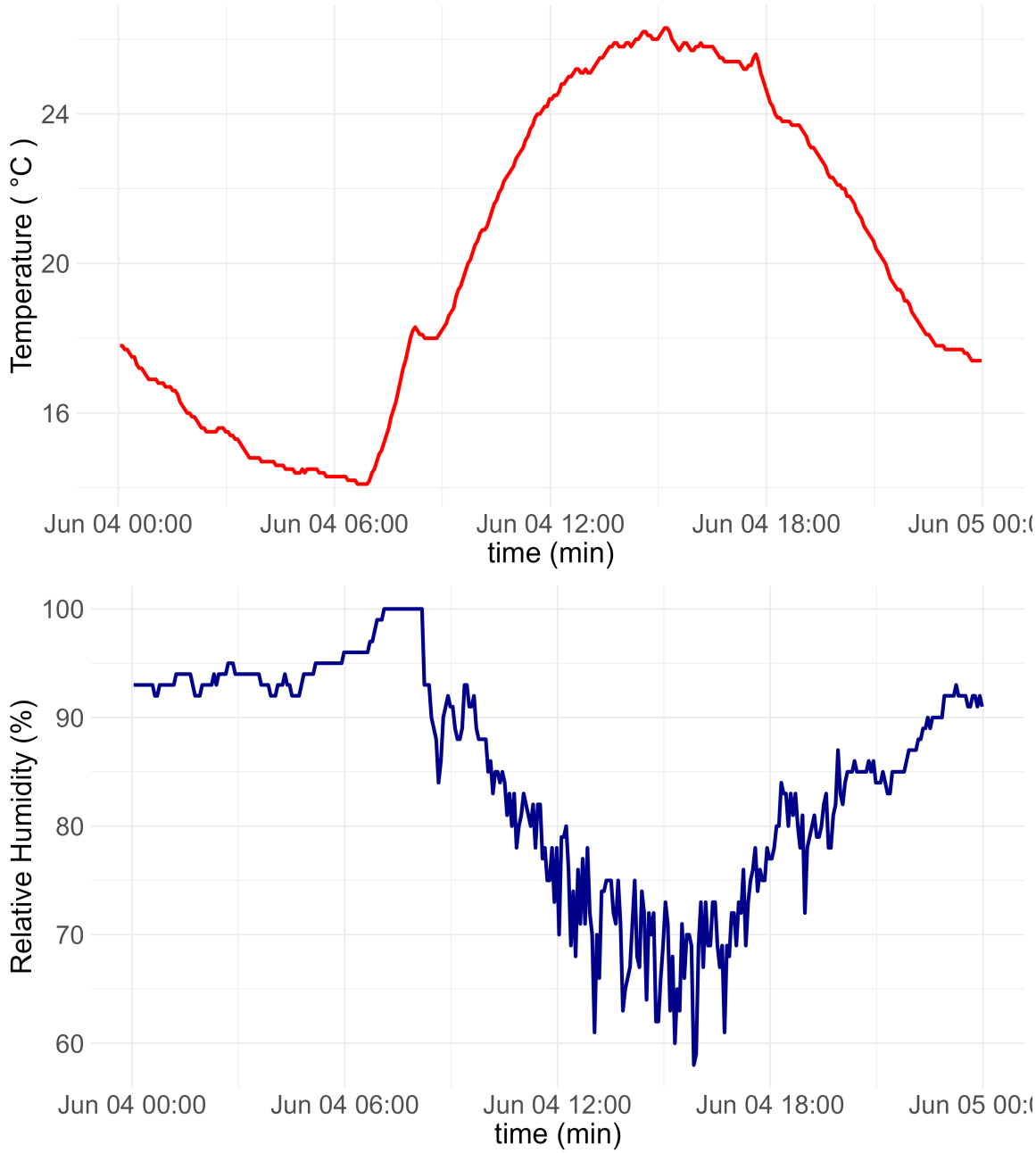


FIGURE 2.3: Minute measurements of Temperature (top), Humidity (bottom) 4/6/21

**Definition 2.3.** *Let $T_b$ represent the base temperature, $T_c$ the ceiling temperature, and $T_{min}(t)$ and $T_{max}(t)$ the minimum and maximum temperature at day $t$. Then,*

*(i) the Growing Degree Days at day $t$, denoted by $GDD(t)$, are defined by*

$$GDD(t) := \frac{T^c_{max}(t) + T^b_{min}(t)}{2} - T_b, \tag{2.1}$$

*where $T^c_{max}(t) = \min\{T_{max}(t), T_c\}$ and $T^b_{min}(t) = \max\{T_{min}(t), T_b\}$,*

*(ii) the Accumulated Growing Degree Days until day $t$, denoted by $AGDD(t)$, are defined by*

$$AGDD(t) := \sum_{s=1}^{t} GDD(s). \tag{2.2}$$

# 4 Humidity

Humidity, a fundamental concept in the study of atmospheric science, pertains to the amount of water vapor present in the air, constituting a crucial factor in determining weather patterns and influencing human comfort levels (Wallace and Hobbs, 2006). Although there exist multiple methods to quantify humidity, the most commonly employed measures are absolute humidity (AH), specific humidity (SH), and relative humidity (RH) (Stull, 2015).

**Definition 2.4.** ***Absolute humidity*** *refers to the mass of water vapor per unit volume of air (g/m³), which fluctuates in response to changes in air temperature and pressure (Figure 2.3).* ***Specific humidity****, on the other hand, is the mass of water vapor per unit mass of moist air (g/kg), remaining invariant with respect to changes in air pressure. Moist air is defined as air containing water vapor, while air composed of a mixture of gases (i.e. oxygen, nitrogen) without considering water vapor content is referred to as dry air.*

*Unlike AH and SH,* ***relative humidity*** *constitutes a dimensionless ratio, expressing the amount of water vapor in the air as a percentage of the maximum capacity at a given temperature and atmospheric pressure.(Stull, 2015).*

Figure 2.3 displays the inverse relationship between temperature and RH. It is worth mentioning that the greenhouse windows were opened a little before 9:00 am, leading to a slight temperature decrease and the release of water vapor, which in turn caused a decline in RH. The windows were closed at 20:00 pm. Fluctuations in the humidity measurements can be attributed to plant transpiration. As moist air exits the greenhouse, plants contribute new moist air, resulting in a cycle that produces these variable measurements (Monteith, 1977).

The significance of humidity in atmospheric processes is paramount, as it not only impacts the formation of clouds, precipitation, and air quality, but also affects plant growth, agricultural

productivity, and the distribution of ecosystems (Trenberth et al., 2015), (Taiz et al., 2015). In the realm of plant growth, humidity plays a critical role in regulating transpiration rates, nutrient uptake, and gas exchange, with both excessively high and low humidity levels capable of adversely affecting plant health and development (Larcher, 2003).

Accurate measurements of humidity are, therefore, essential for improving weather forecasting, facilitating climate modeling, and addressing environmental challenges, as well as optimizing agricultural practices and ensuring the sustainable management of natural resources. By understanding the complex interplay between humidity and various biological, physical, and chemical processes, researchers and practitioners can work towards developing effective strategies to mitigate the impacts of climate change and safeguard our planet's ecosystems.

Difference in values of humidity, create deficits, a concept that will be introduced with Vapor Pressure Deficit (VPD) explained in later section. VPD is the last key driving factor of evapotranspiration.

# 5 Transpiration

Transpiration, a vital physiological process in plants, refers to the movement of water from roots to aerial parts, culminating in the loss of water vapor from plant surfaces, primarily through stomata located on the leaves (Taiz et al., 2015). This process plays a crucial role in maintaining plant water balance, facilitating nutrient uptake, and driving gas exchange for photosynthesis (Larcher, 2003). It is regulated by a combination of biotic and abiotic factors, including plant anatomy and morphology, soil moisture, temperature, humidity, and light intensity (Jones, 1992).

Transpiration can be categorized into two primary types: cuticular and stomatal transpiration. Cuticular transpiration occurs through the waxy cuticle layer covering the leaf surface, constituting a minor component of total water loss due to the relatively low permeability of the cuticle (Schönherr, 2006). Stomatal transpiration, on the other hand, accounts for the majority of water loss in plants and is closely linked to the opening and closing of stomata, which are controlled by guard cells in response to environmental cues and internal signals (Hetherington and Woodward, 2003).

Understanding the dynamics of transpiration is essential for optimizing plant growth, agricultural productivity, and water management. Accurate measurements of transpiration rates, alongside comprehensive knowledge of the factors that influence them, are crucial for developing effective strategies to improve crop performance under changing environmental conditions.

Within this discussion, transpiration will be considered as part of evapotranspiration. However, understanding transpiration is also vital for grasping biological modeling and the link to Water

Consumption, which will be introduced in the following chapter. A better intuition about the physics behind transpiration can be found in the next section.

# 6 Vapor Pressure Deficit (VPD)

Vapor-pressure deficit, or VPD, is the difference (deficit) between the amount of moisture in the air and how much moisture the air can hold when it is saturated. Vapor pressure deficit (VPD) represents a vital parameter that significantly impacts the process of evapotranspiration (ET). In the following section, we shall explore this topic in greater detail. It is essential to note that VPD plays a fundamental role in crop models, as evidenced by various studies in the field (Castellvi et al., 1996).

In greenhouses during the winter season, a high vapor pressure deficit (VPD), particularly due to air exchange during midday period, can limit plant biomass and yield. Thus, the VPD control in greenhouses is of immense importance for cultivating plants (Lu et al., 2015). The VPD model is based on the following two hypotheses:

- Hypothesis I: The VPD is a symmetrical function of time (x-axis) with respect to the y-axis (representing VPD values).
- Hypothesis 2: The actual vapor pressure ($e_a$) remains fairly constant throughout the day.

Under these assumptions, the resulting model is as follows (Castellvi et al., 1996):

$$\begin{aligned} E_{sat} = a \exp\left(\frac{b \cdot T_{air}}{c + T_{air}}\right),\ E_{air} = E_{sat} \cdot RH \\ VPD = E_{sat} - E_{air} \\ = E_{sat}(1 - RH) \end{aligned} \tag{2.3}$$

where, VPD is measured in kPa, $T_{air}$ the ambient Temperature ($^oC$), $E_{sat}$ the saturation vapor pressure (kPa), $E_{air}$ partial pressure of the water vapour in the air (kPa), RH the relative humidity (%) and a,b,c location depended constants. This formation is generally known as the *Magnus form*.

Constants $a, b, c$ are usually approximated. In the agricultural bibliography Buck (1981) investigated various intervals for Temperature values. Castellvi et al. (1996) present a proof of *Magnus form* derivation and compare different methods for estimating VPD, suggesting the 2.3 formulation. In our context, VPD is introduced only to provide the needed intuition for the deficit that drives transpiration and introduce vapor pressure (which will be used in Evapotranspiration).

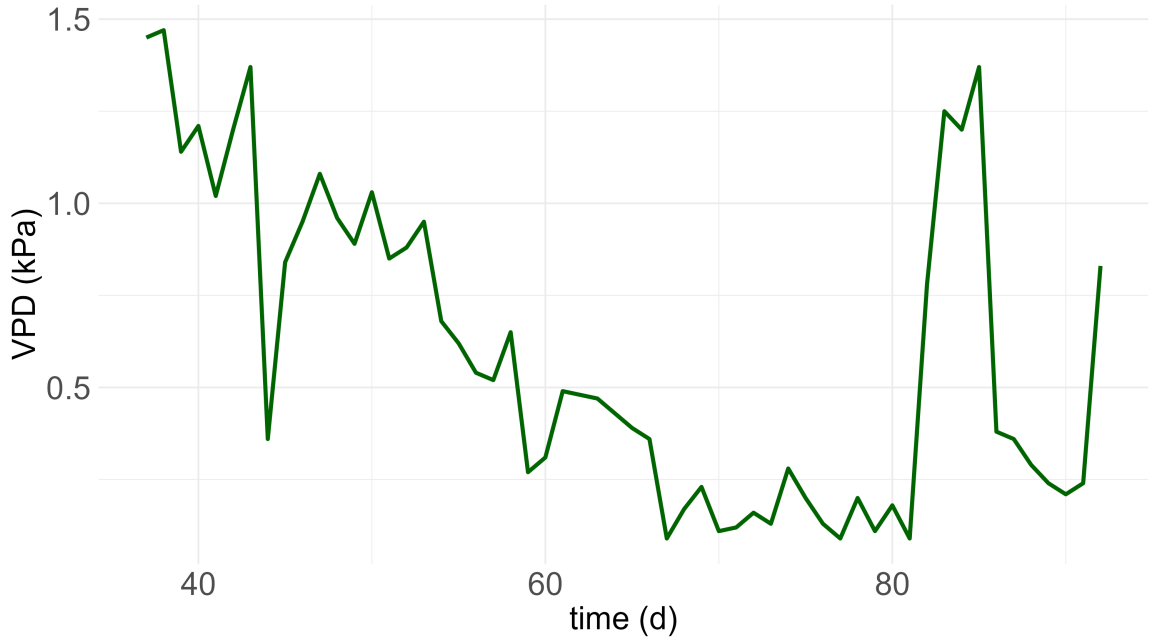


FIGURE 2.4: The figure illustrates the daily average VPD measured during the summer months of 2021.

Intuitively, VPD assists plant water circulation, as the difference between the leaf's water-saturated surface and the air's partially filled water pressure encourages water to exit the leaf. While this explanation does not encompass the entire underlying mechanism, it serves as a step towards a comprehensive understanding.

Visual inspection of VPD (Vapor Pressure Deficit) values can also contribute to our understanding of the phenomenon. To facilitate this examination, we can refer to the graphical representation presented in Figure 2.4 and Figure 2.5. The notable shift observed on day 81 (Figure 2.4) can be attributed to a change in the measurement instrument's position, from a lower to a higher point, leading to lower relative humidity measurements than the ones observed at the lower point. The same shift can be observed at day 84, where the instrument were return to its initial position. This example can highlight the sensitivity of calculated VPD values with respect to RH values, and how much can VPD vary in the same space. Fig 2.5 reveals the extreme change of VPD values with respect to Humidity, under normal stable Temperature values ($\sim 25 - 30^{o}C$).

## 7 Evapotranspiration

Evapotranspiration (ET) is a vital process involving the transfer of water from the surface to the atmosphere via evaporation and plant transpiration. Potential evapotranspiration (PET) is a significant parameter indicating the environmental demand for ET. PET signifies the ET rate for a short, uniformly

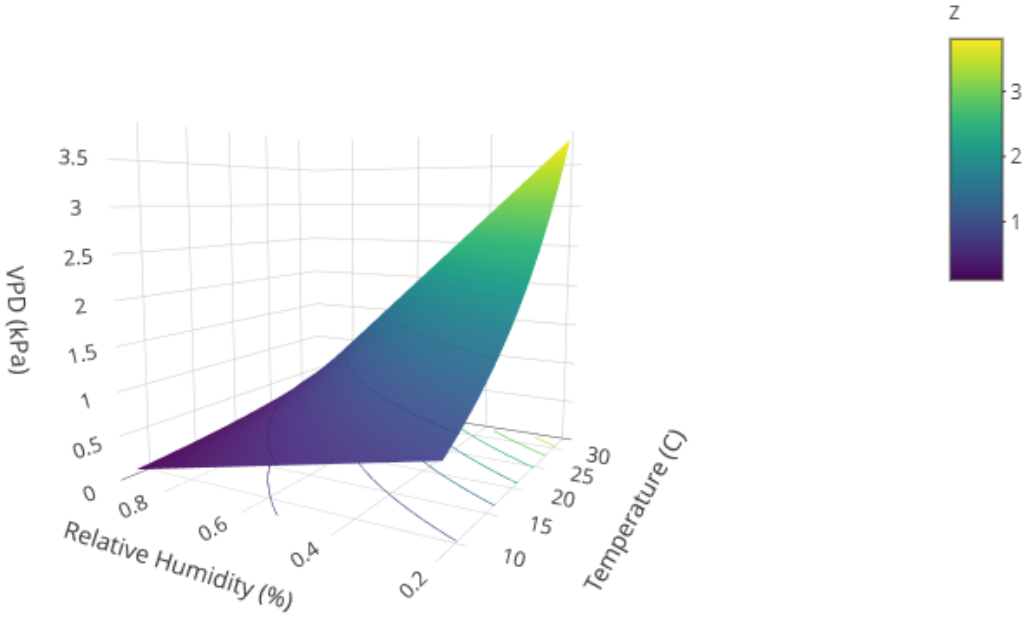


FIGURE 2.5: VPD in 3d

tall green crop with sufficient water availability in the soil. It captures the energy accessible for water evaporation and the wind's capacity to carry water vapor from the ground into the lower atmosphere.

Numerous models are available to estimate PET in agricultural systems, including Penman-Monteith, Thornth-waite, Hamon, Hargreaves-Samani, Turc, Makkink, and Priestley-Taylor. The Food and Agricultural Organization of the United States (FAO) suggests the use of Penman-Monteith. However, it has been shown that estimating PET accurately is challenging, and caution should be exercised when using PET to estimate actual water loss from natural systems (Lu et al., 2005).

In this study, the Priestley-Taylor method was selected for estimating PET for two main reasons. First, unlike the Penman-Monteith method, the Priestley-Taylor method does not necessitate the leaf area index (LAI), a parameter that is challenging for an average producer to obtain. Second, the Priestley-Taylor method has demonstrated good performance and primarily requires a radiation measurement, which is more readily accessible (Lu et al., 2005). The Priestley-Taylor equation for PET estimation is presented below:

$$\lambda PET = \alpha \frac{\Delta}{\Delta + \gamma}(R_n - G) \tag{2.4}$$

where, $PET$ the Potential Evapotranspiration (mm/day), $R_n$, net radiation $(W/m^2)$, in our case Solar irradiance, $\lambda = 2.501 - 0.002361 \cdot T$: latent heat of vaporization (MJ/kg), $T$ average daily temperature ($^oC$), $\alpha = 1.26$ a calibration constant, $\Delta$ the slope of the saturation vapor pressure

temperature curve (kPa/$^oC$), $G$ Soil heat flux density $(W/m^2)$ and $\gamma$ the psychrometric constant (kPa/$^oC$)

Potential evapotranspiration actually combines all the need variables as promised. Calculated PET, over a day period, can be observed in Figure 2.6. Notably, variations in solar radiation contribute to the observed noise in the data. A distinct upward trend is visible from the start until around day 50, followed by a subsequent downward trend. These patterns can be attributed to the Earth's movement and are consistent with the trends observed in Figure 2.1.

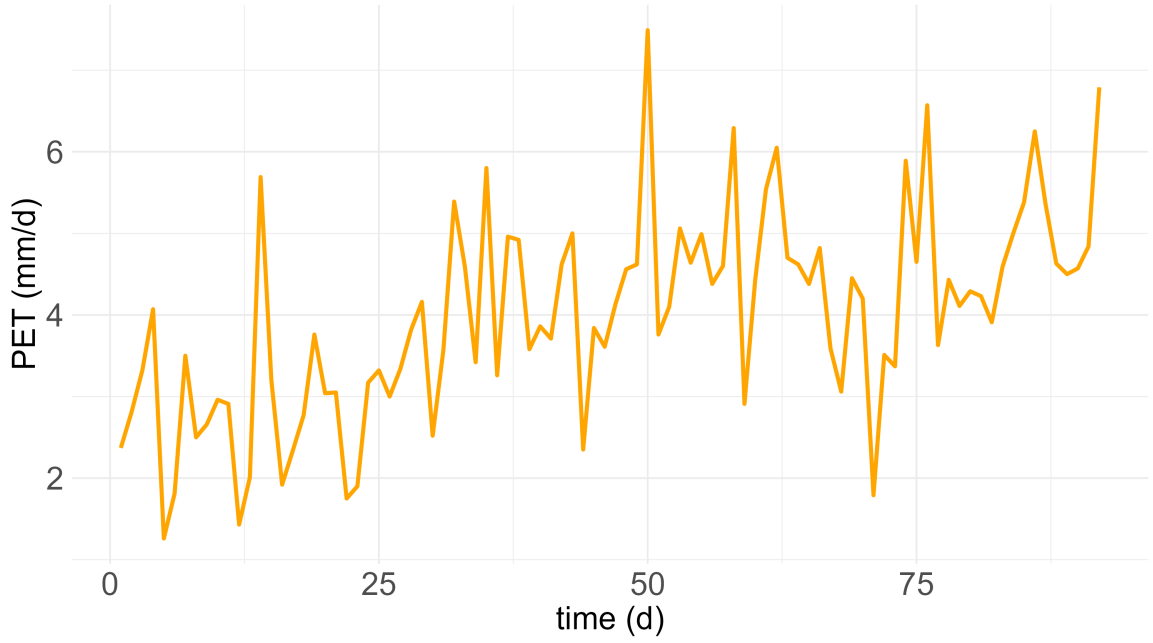


FIGURE 2.6: The potential evapotranspiration (ET) during the summer months of 2021 is estimated using the Priestley-Taylor method.

# Chapter 3

# Knowledge Driven Models (KDM)

## 1 Terminology

In this section we explain some notions which are important for the development of this master thesis and we introduce the necessary notations:

- Irrigation Volume ($IrrV$): the volume of water applied to plants.
- Slab: a substrate on which plants are placed. In the greenhouse where this study was conducted, a slab consists of 3 plants.
- Station: a combined system of a slab and a pot which serves for water collection.
- Runoff ($R_off$): the excess water applied on a station. It is collected in the pot, under the slab.
- Phytomer: a structure comprising an internode that ends in a node on which organs (leaves, fruits and axillary meristems) are attached (De Reffye et al., 2021).
- Cycle of development (CD): the average duration, in thermal time, required to place a new phytomer at the end of the plant main stem is called the cycle of development (CD).
- truss: A collection of fruits attached to the plant, including the petiole which connects them.
- crown: The combination of primary bearing axes and secondary branching axes involves quantifying the phytomers produced per axis in a plant structure. A tree consists of numerous primary and secondary plant canopies. Along the main stem, it is generally the same branched limb that extends until growth ceases, resulting from the termination of the apical meristem.

**Definition 3.1.** *The per average plant volume of water consumed by a station is referred to as* Water Consumption *($W_c$), and it is calculated as follows:*

$$W_c(t) = \frac{V_{Irr}(t) - R_{off}(t)}{n}, \tag{3.1}$$

*where n is the number of plants on the station which refers to a group of plants all growing on the same substrate.* $V_{Irr}(t)$ *is the volume (L) of applied water at time t, and* $R_{off}(t)$ *the volume (L) of the corresponding collected excess water.*

It should be noted that $R_{off}$ is measured the next morning, leaving approximately 12 hours between collection and measurement. In our study, $n = 3$, and the water consumption $W_C$ corresponds to the dependent variable that we try to estimate and predict.

# 2 Models with biological representation

Tackling biological problems, models that can dynamically capture the changes in plant evolution are considered. Examples of this type include the LNAS model (Log-Normal Allocation Senescence) (Chen et al., 2013) and the GreenLab model (Yan et al. (2004), De Reffye et al. (2021)). Despite the lack of biomass measurements, it is challenging to investigate the idea of recovering plant's growth dynamics by measuring only water consumption. The idea behind this approach is that transpiration and biomass production can be considered proportional (Howell and Musick, 1985). Since the GreenLab model will play a central role in the modeling approach, this chapter serves as an introduction to this model and explains how it will be utilized to address the problem at hand. This model will also settle the ground for an innovative approach discussed at the end of the chapter.

## 2.1 GreenLab

The GreenLab model (Yan et al., 2004) is a functional-structural plant model (FSPM), combining both functional and structural description metabolic processes with phytomer-level structures (De Reffye et al., 2021). Breaking the procedure into those fundamental components enables study from the organ up to the macroscopic level. The ecophysiological concepts assumed in crop models (i.e. thermal time, radiation use efficiency (RUE)) assist the model. As an FSPM there are some restrictions; the study must be conducted on the same fixed genotype (clone, variety), with plants of the same age, under the same conditions (i.e. Temperature, light, humidity). Furthermore, GreenLab is a discrete, mathematical model with a limited set of variables and physically interpretable parameters, enabling parameter estimation, model analysis, model evaluation, optimization, and control of farming systems.

For the simulation of the development of the plant, it is sufficient to define the rules regulating the physiological age value of any recently produced phytomer (De Reffye et al., 2021). The next section describes this process.

Organogenesis depends heavily on thermal time, according to a base temperature of $12^oC$ (Shamshiri et al., 2018). A hypothesis done for simplicity is that the thermal time elapsing between the growths

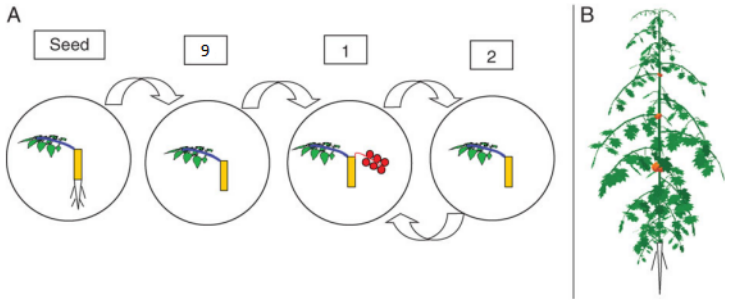


FIGURE 3.1: (A) Greenlab automaton for tomato development and (B) resulting plant architecture at GC 25. (Dong et al., 2008)

of two successive phytomers is constant. This process can also be stochastic, as random events can disrupt the development and architecture of the plants. In the context of this study, it is assumed deterministic, and the stochastic case is left for future developments.

### 2.1.1 Tomato automaton

Tomato's plant construction can be described utilizing four main organs (Dong et al. (2008),Zhang et al. (2009)), including *blade* ("b") (i.e. leaves), *petiole* ("p") (the part that connects the blade to the main stem), *internode* ("e") (the intermediate section that links the end and the start of a phytomer) and *fruit* ("f"). Flowers are assumed fruits from their first inflorescence.

In the usual practice of single-stem, pruned cultivars grown in greenhouses, seven to eleven phytomers without a flower are developed before the first inflorescence. In our case after the seed cycle, the plant is assumed to grow 8 phytomers with one internode, one petiole and one blade. From that stage on and every third phytomer a truss appears. Figure 3.1 illustrates this procedure. Trusses are assumed to generate 3 flowers, that all bud, representing the average number of fruits per crop cycle.

### 2.1.2 Biomass production

Starting from the initial biomass of the seed, denoted $Q(0) = Q_0$, at each time step $t$, the newly accumulated biomass $Q(t)$ will refer to the total biomass produced. A typical relation, derived by the Beer-Lambert law (Monteith, 1977) is:

$$Q(t) = \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-k \frac{Sf(t)}{S_p}\right)\right) \tag{3.2}$$

where $E$ is the environmental parameters (i.e., ET, PAR, or sometimes water), depending on the choice of the growth driving factor, k is analogous to the extinction coefficient of Beer-Lambert's Law. In the case of tomato plants, a value of 0.8 has been proposed (Zhang et al., 2009)) to evade identifiability issues. $S_f$ denotes the total plant leaf surface area, which is a function of the total biomass of the leaves, also referred to as the green biomass of the plant, and the specific leaf area (SLA), a quantity which represents the size of leaf area a plant builds with a given amount of leaf biomass, $S_p$ is the theoretical projection surface of the individual plant and is a parameter under estimation, and RUE the Radiation Use Efficiency (RUE).

### 2.1.3 Allocation

The biomass ascribed to every organ, spread from the common pool, is set proportional to its sink strength. Sink strength adjusts during the period of organ expansion, following the same form of sink function for all organs of the same type $o \in \{b, p, e, f\}$ in a cohort. A cohort is a set of organs of the same nature, created simultaneously by the parallel functioning of meristems.

If $T_o$ stands for the expansion duration of an organ of type $(o)$ and $t$ stands for its chronological age (days or CDs) , then the sink strength is modeled by the function:

$$P_o(t) = p_o \cdot f_o\left(\frac{t}{T_o}\right), \qquad 0 \leq t \leq T_o, \tag{3.3}$$

where $p_o$ is its relative sink strength (with respect to the blade's one), $f_o(\cdot)$ is the variation function of the sink related to its development. The GreenLab model defines the sink function according to a discretized beta law function:

$$f_o\left(\frac{t}{T_o}\right) = \frac{1}{M}\left(\frac{t}{T_o}\right)^{(a_o-1)}\left(1-\frac{t}{T_o}\right)^{(b_o-1)}, \qquad 0 \leq t \leq T_o, \tag{3.4}$$

where parameters $a_o$ and $b_o$, verifying the constrain $a_o, b_o \geq 1$, drive the curve shape, and M is the normalization constant.

The sum of the sink strength of all organs is the Demand $D(t)$ at a given time $t$:

$$D(t) = \sum_o \sum_{u=1}^{t} N_o(t-u+1)P_o(u), \tag{3.5}$$

where $N_o(t-u+1)$ is the total number of organs of type $o$ at time $t$ that appeared at time $u$. The biomass growth of an organ $o$ varies on the value of its sink and the ratio supply produced to the previous cycle $Q(t-1)$ (3.2) by the present demand $D(t)$ (3.5). The expansion of the organ of type $o$

appearing in cycle $u$ when the plant is at cycle $t > u$ is written:

$$\Delta q_o(u,t) = P_o(t-u+1)\frac{Q(t-1)}{D(t)}, \tag{3.6}$$

and the weight of the organ that appeared in cycle $u$ when the plant is at age $t$ is then:

$$Q_o(u,t) = \sum_{j=u}^{t} \Delta q_o(u,j) \tag{3.7}$$

A code that can simulate GreenLab, as described above, in R language, can be found in the Appendix. The code is broken into two parts. The first simulates the automaton of tomato, and the second one brings all the pieces together.

### 2.1.4 Senescence

Pruning is a common agricultural practice which aims to prevent plants from allocating biomass to mature leaves. In some cases, pruning is applied before the senescence of the green biomass. As a result, instead of a function to describe senescence, we introduce a simple function:

$$Q_s(t) = d\,\mathbb{1}_{\{\text{Prune}\}}(t),$$

which subtracts $d$ grams $(g)$ of biomass every time leaves are pruned.

### 2.1.5 Modifications

Various modifications can be applied to the aforementioned modeling, primarily to reduce the parameter space dimensions and/or improve convergence properties. In our context, all the modifications we will utilize involve different versions of the sink function 3.4. Two of these modifications concern the normalization constant $M$, and three of them pertain to the parameters $(a_o)$,$(b_o)$, $o \in \{b,p,e,f\}$. In total, we will explore six distinct modifications.

Normalizing the sink function (3.4) can be done either by the maximum (Zhang et al., 2009) or by the sum (Dong et al., 2008) with respect to $0 \leq t \leq T_o$.

Regarding the parameters $(a_o)$, $(b_o)$, $o \in \{b,p,e,f\}$, the simplest assumption is to treat them as free variables. This approach will be referred to as the *beta sink*. To reduce dimensionality, one constraint can be applied by assuming two distinct $(a_o)$ values for phytomers with $(a_{tr})$ or without truss $(a_{ntr})$ (Zhang et al., 2009). This modification called *beta sink 2*, reduces the parameter space dimension by 2.

Lastly, Dong et al. (2008) proposed the constraint $a_o + b_o = 5$, specifically for tomato plants, which naturally results in a 4-unit reduction in parameter space dimension.

The AIC (Bozdogan, 1987) and BIC (Schwarz, 1978) criteria were utilized to select the best modification.

## 2.2 Link to Water Consumption

Howell and Musick (1985) demonstrated that transpiration and biomass production are proportional. Environmental conditions in one of the experiments they present, discussed in Howell et al. (1984), are similar to the conditions of the experiment under discussion, Table 5.1. In our greenhouse setting, evaporation is assumed to be negligible, so transpiration could in turn be considered proportional to water consumption (Food and Agriculture Organization of the United Nations, 1998) thus rendering the latter linearly related to dry matter production. Disregarding evaporation is not a particularly far-fetched premise within the framework of hydroponic greenhouses. These greenhouses are designed to reduce evaporation to a minimum, utilizing substrates wrapped in white sacks that offer a minimal surface area for water to evaporate from (Resh, 2022).

Adding normally distributed homoskedastic errors, we obtain the following initial model:

$$W_c(t) = \mu_0 \cdot Q(t) + \varepsilon_t, \quad \text{where} \quad \varepsilon_t \sim N\left(0, \sigma^2\right), \tag{3.8}$$

$\mu_0$ is a positive proportionality constant and $\sigma^2$ is a variance parameter.

As $W_c$ measurements were conducted daily, but the GreenLab model runs on Cycles of development (CD), we need to map CDs on days. Elapsed days between two successive leaf developments (phyllochron) can vary from 1.5 (summer) to 3 (autumn) days according to the genotype, and the climatic conditions (Pivetta et al. (2007),Schmidt et al. (2017)). We assume that the phyllochron is stable and equal to 2 days, as we measured a mean value of 10 $^oCd$ with a base temperature of 12$^oC$. To aggregate the two separate measurements into one CD, a weighted average is utilized with a weight proportional to the fraction of the Solar radiation of each day.

### 2.2.1 Log-Likelihood of the model

Given the model we described in the last paragraph, we yield the following log-likelihood:

$$l(\underline{\theta}) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu_0 \cdot Q(t_i))^2}{2\,\sigma^2}$$

for $x_i$ water consumption observations at time $t_i$. For maximization, a bounded version of the quasi-Newton algorithm was utilized (Byrd et al., 1995).

### 2.2.2 A comment on Water Use Efficiency

Generally, agronomists have a unique name for the fraction of Biomass over Transpiration $(Q/T)$. This quantity is called Water Use Efficiency (WUE) in the bibliography. This quantity is known to vary according to environmental conditions, genotype, or even cultural practices (Monteith, 1977). Such a variation can be observed also in Figure 3 of Lanoue et al. (2017). A natural question that arises from these observations is: How such a major proportionality assumption between Transpiration and Biomass production will affect the results of this thesis?

We return to Figure 3 of Lanoue et al. (2017) to investigate this question. From this figure, we can extract information about the WUE and Transpiration profiles. A first observation is that values of WUE, for winter months, range between $5\,\mu molsCO_2/mmolsH_2O$ to $15\,\mu molsCO_2/mmolsH_2O$. This translates to $4 \cdot 10^{-3} g/L$ to $1 \cdot 10^{-2} g/L$. Using the following relation:

$$\begin{aligned} Q &= WUE \cdot T \\ &= \mathbb{E}[WUE] \cdot T + (WUE - \mathbb{E}[WUE]) \cdot T \end{aligned} \tag{3.9}$$

where $Q$ represents the produced biomass in grams $g$, over a $m^2/d$, $T$ the transpiration in $L$ over a $m^2/d$.

From the previous relation, we can deduce:

$$|Q - \mathbb{E}[WUE] \cdot T| = |(WUE - \mathbb{E}[WUE]) \cdot T|$$

As we already discussed, the term in parenthesis, on the right-hand side, can take a maximum value of $0.01 g/L$. On the other hand, the Transpiration, by the same reference, can take a, very loosely chosen maximum value of 24 mmols $H_2O/(m^2 \cdot s)$, which translates to 86.4 moles $H_2O/(m^2 \cdot d)$. To convert this to $L$ we divide by 55.5, an approximation of the number of moles in $1\,L$ of $H_2O$. The result we yield is $1.56\,L/(m^2 \cdot d)$. Returning to equation 3.9:

$$\begin{aligned} |Q - \mathbb{E}[WUE] \cdot T| &\lessapprox 0.01 \cdot 1.56 \\ &= 0.016\, g/(m^2 \cdot d). \end{aligned}$$

Even if we arbitrarily double the maximum value of WUE, to account for the difference in season, this value can go up to $0.032\, g/(m^2 \cdot d)$. As we will see in the forthcoming analysis, this value is by

one (1) order of magnitude smaller than the estimated standard deviation of the model 3.8 for all the different choices of the biomass modeling. The claim we make here is that the fluctuations of WUE among days are incorporated into the error term we already included in the model 3.8.

## 2.3 Identifiability issues and compartmental simplification of the GreenLab model

In our realistic setting where no plant data is available, estimating the parameters of the complete Greenlab, model is unrealistic: identifiability problems will necessarily be encountered. For example, a plant with considerable internodes and negligible petioles could have the same measured Water Consumption as one with reversed features while maintaining equal total biomass.

Adopting a general dimensionality reduction strategy for non-identifiability issues—outlined in (Hastie et al., 2009)—we analyzed a simplified version of the model. We trade precision in representing the biological model for enhanced identifiability within the parameter space. In this version, we combined all the biomass of petioles ($p$), internodes ($e$), and fruits ($f$) into a single representative referred to as *body*.

Parameters requiring estimation thus comprise:

$$\theta = (a_b, b_b, p_{body}, a_{body}, b_{body}, S_p, \Lambda, SLA, \mu_0, \sigma, Q_0) \tag{3.10}$$

We will refer to this specific parametrization as *comp1*.

To explore the identifiability of parameters, we simulate data from the *comp1* model, initialize 5000 starting points, and record the solutions obtained from the minimization of the negative log-likelihood of the model via the similar bounded version of a quasi-Newton algorithm, also used for the maximization of the log-Likelihood (Byrd et al., 1995), with a $10^{-3}$ tolerance threshold.

To present the identifiability issue at hand, we could choose many of the possible subsets of the parameter space to stabilize. If non-identifiability is present in such a context, the general case with the full parameter space is inapproachable (Hastie et al., 2009). For the sake of simplicity, we chose to present two emblematic cases only: in the first one, we set $SLA$, the specific leaf area, and $Q_0$, the initial biomass of the seed, that can typically be measured, along with $S_p$ and $\Lambda$ as we incorporated the $\mu_0$ parameter in the model 3.8. For the second case, we stabilize $P_{body}$, the sink strength of the 'body' compartment. All stabilized values are chosen to be the ones of their respective 'true' parameter value. Our rationale behind the choice of those cases is their relevance in a realistic setting. The first one can be practically applied, as both $SLA$ and $Q_0$ can be measured. The second one is practically irrelevant, as there is no procedure to measure $P_{body}$. All stabilized values are chosen to be the ones of their respective 'true' parameter value.

Separating those two cases can give us insights into this methodology's applicability level. If the first case is feasible, then the results obtained from it could potentially be relevant, whereas if not the method is infeasible. In the second case, positive results of the test could showcase the ability to identify some of the parameters, under some heavy assumptions.

## 2.4 Two model versions for water consumption timeseries based on the recurrence equation of GreenLab

As shown in Letort et al. (2009), the GreenLab model can be synthesized into a single recurrence equation that, for the sake of simplicity, we chose here to formulate as:

$$Q(t) = \text{RUE} \cdot E(t) \cdot S_p \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{n=0}^{t-1} r(n) Q(n)\right)\right),$$

where $r(n)$ represents the proportion of green biomass from the totally produced biomass $Q(n)$, quantities that, in the case of the GreenLab model, are functions of the parameters of the model. To show this claim, consider this: by the BL law 3.2 and the total biomass allocated to blades, equation 3.7, we have:

$$\begin{aligned} Q(t) &= \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-k \frac{Sf(t)}{S_p}\right)\right) \\ &= \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{n=1}^{t} Q_b(n,t)\right)\right) \\ &= \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{n=1}^{t} \sum_{u=n}^{t} \Delta_b(n,u)\right)\right). \end{aligned}$$

In the last equation, we also made use of the assumption that at each CD, we only have only one leaf for these tomato plants under this pruning strategy. If this did not hold, we would need to include the total number of leaves that appeared on the same CD. We now substitute 3.6 to the final equation to get:

$$Q(t) = \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{n=1}^{t} \sum_{u=n}^{t} P_b(u-n+1) \frac{Q(u-1)}{D(u)}\right)\right).$$

Now, we change the order of the sums:

$$Q(t) = \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{u=1}^{t} \sum_{n=1}^{u} P_b(u-n+1) \frac{Q(u-1)}{D(u)}\right)\right)$$

and set $r(u-1) = \sum_{n=1}^{u} \frac{P_b(u-n+1)}{D(u)}$, for $u = 1, \ldots, t$. By properly adjusting the index, we yield the claim:

$$Q(t) = \text{RUE} \cdot E(t) \cdot S_p \cdot \left(1 - \exp\left(-\frac{k \cdot SLA}{S_p} \sum_{n=0}^{t-1} r(n)\, Q(n)\right)\right).$$

For simplicity, we avoided to include senescence, but the analysis would have been similar.

Assuming proportionality (with constant $\mu_0$) between biomass production and water consumption and no leaf senescence, we obtain a general model form for water consumption:

$$W_c(t) = \theta_1 \cdot E(t) \cdot \left(1 - \exp\left\{-\theta_2 \sum_{n=0}^{t-1} r(n) W_c(n)\right\}\right), \tag{3.11}$$

where $\theta_1 = \text{RUE}\, S_p\, \mu_0$ and $\theta_2 = \frac{k \cdot SLA}{\mu_0 \cdot S_p}$ are estimated. This model will be referred to as *GreenLab exp*.

To account for the obviously existing differences between the tomato plants in Dong et al. (2008) and our plants, we extend this version by introducing a parametric model of the series

$$r(t) = \frac{t^a}{I(a)}, \quad \text{where} \quad I(a) = \int_0^{t_{max}} t^a\, dt = \frac{t_{max}^{a+1}}{a+1}$$

corresponds to a normalization constant with respect to $a$, a parameter under estimation, and the maximum time of observation $t_{max}$, derived by the experimental design. This model will be referred to as *exp + rate*.

# 3 Stochastic Segmentation of input Energy models (SSiE)

Building upon the prior discussion, we now focus on a novel aspect that broadens the model formulation. Here, we aim to represent biomass production at time $t$, as the cumulative byproduct of a composite stochastic experiment, which consists of many independent individual experiments, each one deciding whether elementary radiative inputs will be absorbed by the plant or not. We thus derive a family of models, which we name 'Stochastic Segmentation of input Energy' models (SSiE).

## 3.1 Formulation of the water consumption series from a stochastic model of light interception

In this section, we discuss the intuition behind a probabilistic interpretation of biomass production, and we formalize this intuition with tools from theoretical probability. Recall that at each time $t$, a total radiative input $E(t)$ is channeled into the system per $m^2$. Assume that this input is equally quantized into very small elementary quantities $\{E_i(t)\}_{i=1}^{n}$ in such a way that either they are completely absorbed

by the plant and converted into biomass by the enlightened parts of the plant or they exit the system without affecting it. In this case, $E_i(t) = E(t)/n$ where $n$ represents the number of "elementary" units. If no other specific details are known, one could assume that the individual events of absorption, say $A_i(t)$, are independent with identical probability of occurrence $p(t)$. With this interpretation and if $\mathbb{1}_{A_i(t)}$ stands for the indicator function of the corresponding event, each elementary radiative input $E_i(t)$ is associated with a random variable

$$Q_i(t) = \text{RUE} \cdot S_p \cdot E_i(t) \cdot \mathbb{1}_{A_i(t)}, \tag{3.12}$$

which records its produced biomass, either 0 if the event $A_i(t)$ is not realised, either RUE $S_p E_i(t)$ if the event is realised, and thus it is totally transformed. The total biomass produced by the plant at time $t$ can thus be expressed as follows:

$$Q^{(n)}(t) = \sum_{i=1}^{n} Q_i(t) = \text{RUE}\, S_p\, E(t) \frac{\sum_{i=1}^{n} \mathbb{1}_{A_i(t)}}{n}. \tag{3.13}$$

Clearly, the last factor of the above expression corresponds to the sample mean of independent and identically distributed random variables and in particular Bernoulli random variables with common probability $p(t)$. Intuitively, one should expect by the strong law of large numbers that the sample mean value should be very near to their common probability of absorption, that is $p(t)$. These arguments give an intuitive interpretation of the fact that the following approximations should be plausible:

$$Q(t) \approx Q^{(n)}(t) \approx \text{RUE}\, S_p E(t)\, p(t). \tag{3.14}$$

However, despite the seemingly sound arguments underlying these approximations, a theoretical justification of their validity is more complex. An obvious theoretical caveat regarding the validity of these approximations is that we cannot conceptualize a countably infinite sequence of events of common probability that play the role of the elementary events of biomass absorption, or equivalently the total radiative input cannot be partitioned into a countably infinite number of positive parts potentially transformed into biomass. The only possibility for justifying the above approximations would be to resort to an uncountable number of stochastic experiments. This approach could still be intuitive but surely involves more mathematical intricacies.

Let us now try to justify the rationale. The radiative input $E(t)$ could be mapped to the interval $[0, E(t)]$ representing an uncountable number of points potentially available for biomass production. At each point $u$ of the interval, one could attach a Bernoulli random variable, say $X_u(t)$, deciding whether the point $u$ will enter the system or not. If it enters the system, then it brings an infinitesimal

contribution to biomass production; otherwise, it is rejected and exits the system. One could still keep the independence assumption and assume that there is a common probability $p(t)$ of the radiative points entering the system, but there is a price to pay. If we assume that the radiative input is a realization of the stochastic process $\{Z_u(t)\}_{u=0}^{E(t)}$, where the sample (observed) paths would be an interval of points consisting of $0's$ and $1's$, then it can be proved with tools from probability theory that the resulting processes are not measurable. To give an interpretation of this nonmeasurability concept, it roughly corresponds to the idea that it would be impossible to associate the usual notion of length to the set of points that entered the system and the set of points that exited the system in this ideally conceptualized experiment. Luckily enough, there is still a solution, and it gives a formal justification for our intuitive approximations. It resides in the disintegration theorem (Chang and Pollard, 1997), a result of measure and probability theory. In fact, this theorem gives very powerful tools and a more intuitive approach to the definition of conditional probability and conditional expectation than the one that is usually presented in standard probability textbooks. A formal description of this theorem and related conditions for its validity would be out of the scope of this paper, and we refer to Chang and Pollard (1997). However, we describe the basic ingredients and the result we need in our context.

Instead of selecting points from the interval $[0, E(t)]$, one could think that the same interval is actually a bundle of Bernoulli experiments, where each one of them is realised when the point $u$ is "activated". Formally, one needs a measure space which consists of the set $Y_t := [0, E(t)] \times \{0, 1\}$, an appropriate measure $\mu$ and a function $\pi : Y_t \to [0, E(t)]$ (usually the projection function) which disintegrates the measure $\mu$ into a family of measures $\{\mu_u\}_{0 \le u \le E(t)}$, such that for a measurable $A$

$$\mu_u(A) = \mu_u\Big(A \cap \big(\{(u, 0), (u, 1)\}\big)\Big) \tag{3.15}$$

and induces the measure $\nu = \mu \circ \pi^{-1}$ on $[0, E(t)]$. In our case, the choices are rather simple. Each $\mu_u$ is "living" (has its support) on the fiber $\{u\} \times \{0, 1\}$ and behaves as a Bernoulli measure, while the induced measure $\nu$ should be the Lebesgue measure restricted on $[0, E(t)]$. In this way, the disintegration theorem justifies the following way of computing the measure of a measurable set $A$:

$$\mu(A) = \int_0^{E(t)} \mu_u(A)\, du, \tag{3.16}$$

where $\mu_u(A)$ is given by (3.15) and the integral should be understood in the Lebesgue sense. We are now ready to make the correspondence with the computation of the totally produced biomass at time $t$. Since the set $B = [0, E(t)] \times \{1\}$ corresponds to the set of all active points, in order to assess the

totally absorbed radiative input, we just have to compute

$$\mu(B) = \int_0^{E(t)} \mu_u(B)\, du = E(t)p(t), \tag{3.17}$$

since $B \cap \{(u,0),(u,1)\} = \{(u,1)\}$ and $\mu_u(\{(u,1)\}) = \mathbb{P}(A_u(t)) = p(t)$. Multiplying by RUE $S_p$ to transform into biomass we get the expected approximation result given by (3.14). It is also interesting to notice that the constant probability $p(t)$ is actually playing the role of a constant flow (with respect to the incoming radiation) of biomass production.

The next step is to appropriately model the probability of absorption $p(t)$, which can classically be done through a parametric family of continuous distribution functions. For each time $t$, let $\{Z_u(t)\}_{u=0}^{E(t)}$ represent the Bernoulli experiments of absorption of the radiative input for all possible $u$ ranging from 0 to $E(t)$. If we denote by $LIS(t)$ the Light Interception Surface at time $t$, then, assuming that the maximum available soil surface is $S_p$, one could construct a new family of random variables $\{U_u(t)\}_{u=0}^{E(t)}$ uniformly distributed on $[0, S_p]$ which concretize the above experiments. In particular, the interval $[0, S_p]$ is partitioned into two subintervals $[0, LIS(t)]$ and $(LIS(t), S_p]$. Then, the absorption events can be written as

$$A_u(t) := \{Z_u(t) = 1\} = \{U_u(t) \leq LIS(t)\}, \quad 0 \leq u \leq E(t). \tag{3.18}$$

In probability theory, such a family exists; loosely speaking, this reinterpretation of the absorption events corresponds to a collection of idealized experiments where an elementary radiative input enters the system if it intersects with the green part of the plant. Now, notice that $p(t)$ corresponds exactly to the probability of the event given by (3.18), which is related to the Light Interception Surface $LIS(t)$ at time $t$. However, $LIS(t)$ is not directly observable, but only indirectly via the cumulated water consumption prior to time $t$, denoted by $SW_c(t^-)$ (itself proportional to the cumulated produced biomass). A novelty of this study consists in making a link between $LIS(t)$ and $SW_c(t^-)$ through an increasing (non-decreasing) function $g : \mathbb{R}_+ \to \mathbb{R}_+$, that is, $LIS(t) = g(SW_c(t^-))$. By the above argument, Eq. (3.18) and the fact that $U_u(t) \sim \text{Unif}(0, S_p)$ we get that all the following equalities hold:

$$p(t) = \mathbb{P}\big(U_u(t) \leq LIS(t)\big) = \mathbb{P}(U_u(t) \leq g(SW_c(t^-))) = \frac{g(SW_c(t^-))}{S_p} = \frac{LIS(t)}{S_p} =: LIR(t),$$

where the last term stands for the Light Interception Ratio. Now, also notice that if $U \sim \text{Unif}(0, S_p)$ is a copy from the family $\{U_u(t)\}_{u=0}^{E(t)}$ and $g$ is invertible, then the third term above can be rewritten as

$$LIR(t) = \mathbb{P}\big(g^{-1}(U) \leq SW_c(t^-)\big) = \mathbb{P}\big(X \leq SW_c(t^-)\big) = F_X(SW_c(t^-)), \tag{3.19}$$

where we set $X = g^{-1}(U)$. In fact, since $g$ is assumed to be an increasing function, its inverse exists at least in a generalized form (generalized inverse), and the above equations still hold. The problem is then to define the relationship between $LIR(t)$ (or $LIS(t)$) and $SW_c(t^-)$ without having any information on the plant itself, and in the next section we discuss several such possibilities.

## 3.2 Different options for the distribution of X

The determination of a mechanistic functional relationship between $LIR(t)$ and $SW_c(t^-)$ is unrealistic. Biologically speaking, the underlying processes are complex and involve, among others, the patterns of biomass allocation to blades and their arrangement in space. An approach to this objective is, however, feasible and a selected number of possible distribution families could be used to compete for their fitting quality and their predictive ability. By introducing additive errors as in Section 2.2, we can derive a model directly applicable to the Water Consumption variable

$$W_c(t) \sim N\Big(\theta_1 \cdot E(t) \cdot F_X\big(SW_c(t^-)\big),\, \sigma^2\Big), \tag{3.20}$$

thereby eliminating the requirement for biomass as an intermediary variable. Each model is determined by specifying $F_X$ in one of the following parametric family of distributions.

**Exponential distribution.** The exponential distribution is one of the most fundamental suppositions that one can make when faced with an undetermined distribution since it corresponds to the maximum entropy solution for a given expected value on the positive line (Jaynes, 1957). Besides, in our setting, it leads to a Beer-Lambert-like model. By (3.19) and the assumption of an exponential model we get:

$$LIR(t;k) = 1 - \exp\big(-k \cdot SW_c(t^-)\big), \quad t \geq 0. \tag{3.21}$$

**Gamma distribution** The gamma distribution is a generalization of the exponential distribution. This provides a logical progression from our initial assumption of an exponential distribution. By (3.19) and the assumption of a gamma model, we get:

$$LIR(t;k,a_\gamma) = \int_0^{SW_c(t^-)} \frac{k^{a_\gamma}}{\Gamma(a_\gamma)}\, s^{a_\gamma - 1}\, e^{-k\cdot s}\, ds, \quad t \geq 0. \tag{3.22}$$

**Mittag-Leffler distribution**

Mittag-Leffler introduced the function bearing his name in 1903 (Bateman, 1953). Different properties of the distribution generated by the Mittag-Leffler function were explored in Pillai

(1990). The concept of generalizing the Beer-Lambert law with the use of the Mittag-Leffler function was proposed by Casasanta and Garra (2018). Following their work, we incorporate this generalization into our analysis, leading to the following LIR term:

$$LIR(t; k, a_{ML}) = 1 - E_{a_{ML}}\big(-(k \cdot SW_c(t^-))^{a_{ML}}\big), \quad t \geq 0, \tag{3.23}$$

where $E_{a_{ML}}$ is the Mittag-Leffler function:

$$E_{a_{ML}}(x) = \sum_{j=0}^{\infty} \frac{x^j}{\Gamma(j \cdot a_{ML} + 1)}, \quad x \in \mathbb{R}, \tag{3.24}$$

with $a_{ML} \in (0, 1]$ the *tail* parameter and $k > 0$ the *rate* parameter. For $a_{ML} = 1$ the above formulation reduces to the exponential distribution with rate parameter $k$.

**Log-normal distribution**

The log-normal distribution is commonly employed to model growth rates Our reasoning for incorporating this distribution in our analysis stems from the presumption that the elementary events $(A_i)_{i=1}^n$ are influenced by the incremental growth of smaller plant elements. This growth is contingent on their size. For the density function, we proceed by adopting the ensuing parametrization:

$$LIR(t; \mu_{\log}, \sigma_{\log}) = \int_0^{SW_c(t^-)} \frac{1}{s \cdot \sigma_{\log} \cdot \sqrt{2\pi}} \exp\left(-\frac{\big(\log(s) - \mu_{\log}\big)^2}{2\sigma_{\log}^2}\right) ds, \quad t > 0. \tag{3.25}$$

**Pareto distribution** The last distribution we explore is Pareto. Following (Van der Zande et al., 2010) (mainly the results depicted in Figures 2 and 3), we observe that the percentage of the biomass responsible for most of the energy interception follows a similar law to the Pareto 80/20 rule (Juran and De Feo, 2010). The formulation of the distribution function that we adopt is as follows:

$$LIR(t; \theta, \eta) = 1 - \left(\frac{\eta}{SW_c(t^-)}\right)^{\theta}, \quad SW_c(t^-) > \eta. \tag{3.26}$$

# Chapter 4

# Results

Between May 10 and July 2, 2021, an extensive study was conducted in a hydroponic greenhouse near Therma village, within the Nigrita-Serres region (40.91, 23.55), Greece, to examine the tomato plant's (cv. ecstasis) water consumption patterns. A drip irrigation system was used to ensure precise irrigation for each individual plant. Rockwool, a product of basalt, was used as a substrate-growing medium in keeping with common practices in the region. Plants' density is reported as 5 stems per $m^2$ (one stem per plant).

Indoor measurements were performed using an Efento Logger. Additionally, meteorological data were collected using a Davis Vantage Pro 2 (Plus) weather station close to the greenhouse. Moreover, daily runoff measurements were conducted and subsequently converted into water consumption data following the methodology outlined at the beginning of Chapter 3. A comprehensive overview of the measured quantities, including Solar Radiation, Temperature, Humidity, and Air pressure, averaged on a daily level, is presented in Table 5.1 ($N = 54$). This chapter presents the findings of this research and provides an analysis of the obtained results.

It is essential to note that the potential evapotranspiration ($ET_{pot}$) was not directly measured during the course of the study. Instead, $ET_{pot}$ was calculated using the method (2.4) described in Chapter 2. This chapter focuses on the results obtained from the analysis of the data collected during the study.

## 1 Linear models

Results of two selection procedures (backward elimination and forward selection), applied on the whole dataset, are presented in table 4.1. The BIC criterion determines each step of the stepwises processes (2.2). Solar radiation measurements have been standardized to prevent influence as a result of magnitude. The resulting BIC values are reported as **-47.62** and **-32.11** respectively. Both the BIC criterion and the $R^2$-adjusted indicate that the model chosen by the backward elimination procedure should be preferred.

TABLE 4.1: Estimated linear models of Water Consumption. Each step of the stepwise procedure is determined by the BIC criterion. The left column presents the final model selected by the backward elimination procedure, while the resulting model of the forward selection procedure is presented in the second column. Solar radiation has been standardized to avoid influence as a result of magnitude.

| | $W_c\ (L)$ | |
|---|---|---|
| | **(backward elimination)** | **(forward selection)** |
| Max ET $(mm)$ | $0.032^{*}$ | |
| | (0.016) | |
| Thermal time $(^oCd)$ | $2.872^{***}$ | |
| | (0.720) | |
| Avg Temp $(^oC)$ | $-2.654^{***}$ | |
| | (0.712) | |
| Max Temp $(^oC)$ | $-0.162^{***}$ | |
| | (0.021) | |
| Max Hum $(\%)$ | $19.251^{**}$ | |
| | (7.724) | |
| Avg VPD $(kPa)$ | $0.219^{**}$ | |
| | (0.096) | |
| Min VPD $(kPa)$ | $10.264^{**}$ | |
| | (4.788) | |
| Cut leaves (Indicator) | $-0.162^{**}$ | |
| | (0.067) | |
| past wc $(L)$ | $0.265^{***}$ | $0.515^{***}$ |
| | (0.085) | (0.100) |
| Solar Radiation (stand.) | $0.737^{***}$ | $0.724^{***}$ |
| | (0.160) | (0.159) |
| time (t) | | $0.014^{***}$ |
| | | (0.003) |
| Constant $(L)$ | $15.008^{**}$ | $-0.491^{***}$ |
| | (5.620) | (0.141) |
| Observations | 53 | 53 |
| $R^2$ | 0.963 | 0.915 |
| Adjusted $R^2$ | 0.954 | 0.909 |
| Residual Std. Error | 0.108 (df = 40) | 0.152 (df = 47) |
| F Statistic | $104.834^{***}$ (df = 10; 40) | $167.652^{***}$ (df = 3; 47) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

TABLE 4.2: Reported Root Mean Square Prediction Error (RMSPE) of the two stepwise procedures under investigation.

| RMSPE | |
|---|---|
| (**backward elimination**) | (**forward selection**) |
| 0.137 | 0.165 |

In terms of forecasting, a sequential methodology is employed. From the original dataset, we initially extract the first 55% days (days 39 to 68) for training and predict the next day's water consumption (day 69). Subsequently, we increase the size of the training set by one additional day at each step, continuing to predict the following day until we reach the end of the time series. The parameters are re-estimated at each step of the procedure. Root Mean Square Prediction Error is used for the comparison of the two models. In table 4.2, the RMSPE error is reported on the two models. Again, the model constructed by the backward elimination process seems to be preferred.

Even though all the criteria we examined suggest that the model derived by the backward elimination method behaves in a more proper manner, we have to comment that all the values used for the prediction are the measured ones. For most of them, except maybe *past_wc* and *Cut_leaves*, an additional predictive model should be incorporated to use such values in a practical predictive setting. In light of this observation, and considering that such hierarchical models would introduce more variability to the prediction, we suggest, between these two, the second model, the one derived by the forward selection method, as the only variable of that model which needs caution in a predictive setting is the Solar radiation variable. This model combines simplicity and predictive accuracy.

## 2 Validating of GreenLab function

To ensure the validity of the results and maintain the integrity of the analysis, the GreenLab function, as described in the Appendix, was fitted to an already researched dataset by minimizing the following loss function:

$$L(s;x) = \sum_{o\in\{b,p,e,f\}} \frac{\sqrt{\frac{\sum(s_i-x_i)^2}{n}}}{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

where, $x = (x_i)_{1:n}$ are the observed values, and $s = (s_i)_{1:n}$ the simulated ones. The dataset used for this fitting process was previously published in (Zhang et al., 2009). The results of this fitting can be found in Figure 4.1.

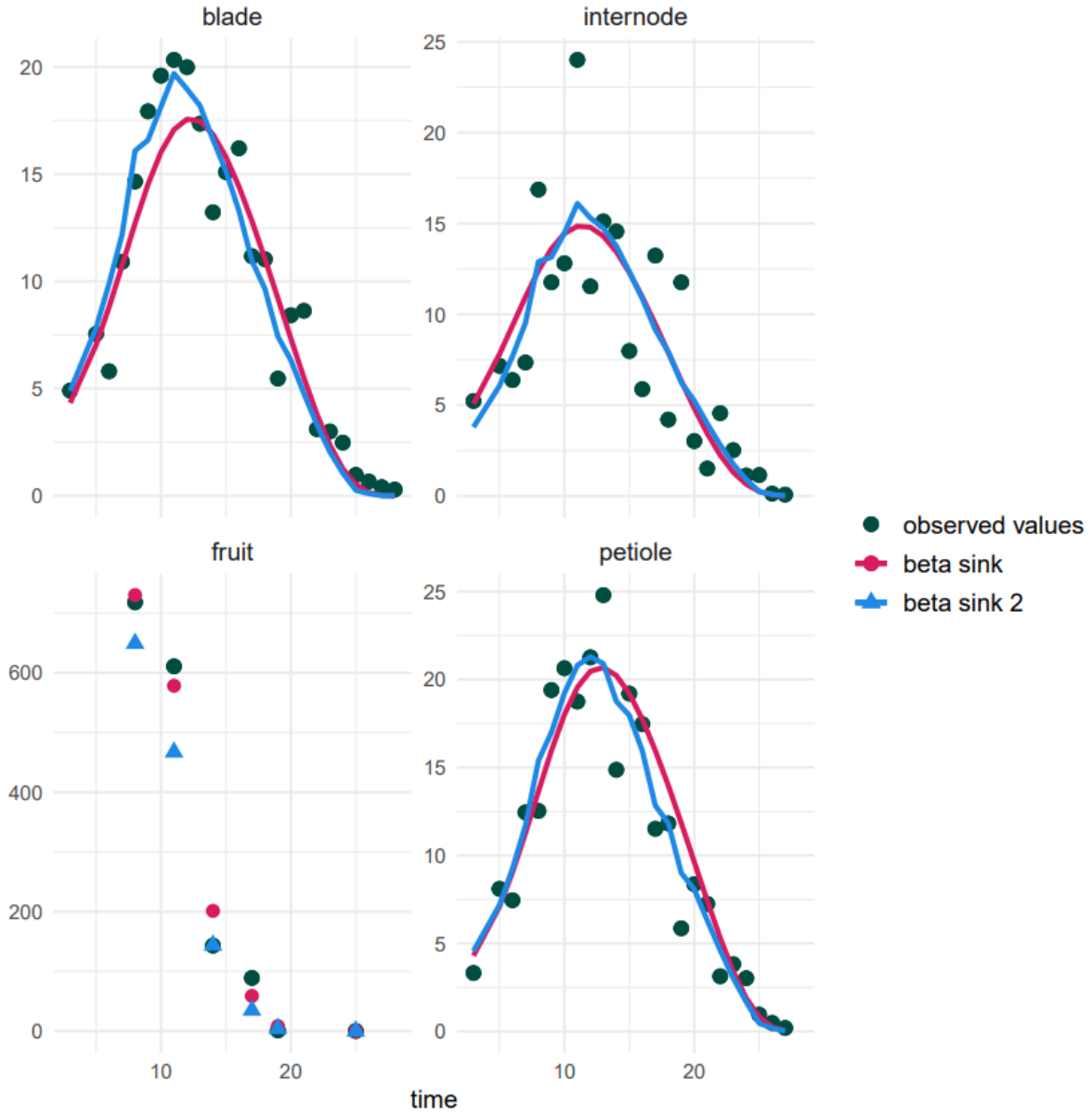


FIGURE 4.1: GreenLab model fit applied to distinct phytomers using the dataset from (Zhang et al., 2009). Data points represent the observed values, while the red and blue lines correspond to the model fit for the two different parameterizations (Beta sink and Beta sink 2, respectively).

The Beta sink and Beta sink 2 functions represent distinct parameterizations of the Beta sink function, as described in Equation 3.4. The primary difference between these two parameterizations lies in the estimation of $a_o$ and $b_o$, $o \in \{b, p, e, f\}$. The parameters of Beta sink function, $a_o$ and $b_o$ are estimated individually for each organ. Conversely, the Beta sink 2 function employs two distinct types of $a_o$ parameters, differentiating between phytomers with and without truss.

An additional contrast between the two parameterizations should be emphasized. As evident in Fig 1 of (Zhang et al., 2009), the Beta sink 2 parameterization generates spikes that provide a

more accurate fit to the data. However, in our study, this improved fitting is not observed. The discrepancy arises due to our lack of access to the environmental data from the original publication, which necessitates the assumption of a constant ET value of 1 for all observations, an assumption that can be though of as constant environmental conditions along the evolution of the phenomenon.

## 3 GreenLab model

In this section, we present the findings of our investigation into the various methodologies outlined in Sections 2.1.5 and 2.3. A comparison using the criteria of AIC and BIC is provided in Table 4.3. Notably, the sink method with '$a_o + b_o = 5$' appears to be the most effective, primarily due to its reduced parameter space dimension and the penalties applied by both criteria proportional to the dimension. These methods seem to under-perform in contrast with the results presented in section 1 of the current Chapter.

TABLE 4.3: Comparison results over different simplification of the GreenLab model.

| | normalization method | sink method | likelihood value | df | AIC | BIC |
|---|---|---|---|---|---|---|
| 1 | max | a_o b_o free | 14.268 | 18 | 30.684 | 66.486 |
| 2 | | a_o+b_o=5 | 11.465 | 14 | 23.121 | 50.967 |
| 3 | | beta sink 2 | 14.207 | 16 | 26.693 | 58.516 |
| 4 | sum | a_o b_o free | 13.698 | 18 | 30.766 | 66.567 |
| 5 | | a_o+b_o=5 | 11.984 | 14 | **23.033** | **50.879** |
| 6 | | beta sink 2 | 14.702 | 16 | 26.624 | 58.448 |

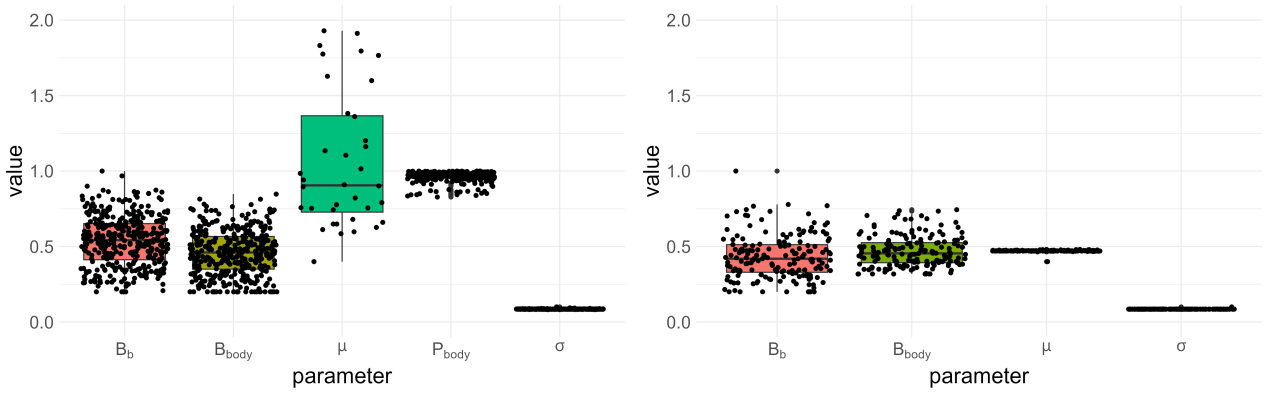


FIGURE 4.2: Boxplots on solutions with similar likelihood values for two stabilization cases and each combination of dots a solution. Each dot represents a parameter value. The sink strength of the body compartment ($P_{body}$) is normalized by its maximum for scaling reasons. (Left) Stabilized parameters: $SLA$, $S_p$, RUE, $Q_0$. (Right) Stabilized parameters: $SLA$, $S_p$, RUE, $Q_0$, $P_{body}$.

Another aspect we already discussed in section 2.3, refers to the identifiability of the parameters. Following the procedure described in the aforementioned section, we get the results presented in Figure 4.2. This figure features box-plots of solutions with proximate likelihood values (distance $< 10^{-3}$), for two test cases: (i) in the first one, all parameters are set to their reference values except $B_b$, $B_{body}$, $\mu_0$, $P_{body}$, and $\sigma$ that are estimated; (ii) in the second one, the parameter $P_{body}$ is also set. Each point represents an estimated parameter value, and specific combinations of these points correspond to the estimated solutions of the maximization problem. For scaling purposes, $P_{body}$ has been normalized by its maximum value. The plots reveal that many distinct solutions yield the same likelihood value. As can be seen by comparing the variation ranges of the estimated parameters between Figure 4.2(left) and 4.2 (right), the implications of this issue diminish as we stabilize more parameters but never disappear. Even with only four estimated parameters, we observe compensation effects between *Bb* and *Bbody*, as their estimation vary. These results indicate that this modeling approach does not produce satisfactory results under our current dataset and methodology, so we do not proceed to a prediction evaluation.

However, a noteworthy outcome of this analysis is that the parameters $\mu_0$ and $\sigma$ are identifiable, at least locally, around the chosen reference values. This observation is significant as these particular parameters also find utility in the stochastic framework elaborated upon in section 3.

# 4 SSiE

## 4.1 Estimation

The outcomes derived from the estimation of SSiE models' parameters are detailed in Tables 4.4 and 4.5. It can be observed that the *lognormal* and *pareto* models demonstrate superior performance in terms of both the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). A straightforward application of Equation 3.11 by introducing estimation of the green biomass by an already fitted model (Dong et al., 2008), however, does not appear to be highly promising, as it results in lower values in these criteria. Similar behavior is present in the Beer-Lambert-like approach of the exponential distribution 3.21. A notable result is the estimation of $a_{ML} \simeq 0.5$, as can be observed in table 4.4. For $a_{ML} = 0.5$ the Mittag-Leffler function (3.24) reduces to (Haubold et al., 2011):

$$E_{1/2}(x) = e^{x^2} \left(1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2}\, ds\right),$$

where $\frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2}\, ds$, also known as the Gauss error function, is a quantity which expresses the probability of a typical Gaussian distribution to be found in the interval $[-x, x]$ for $x \geq 0$. In our case,

it translates to:

$$LIR(t) \simeq 1 - \exp\left(-k\,SW_c(t^-)\right)\ \mathbb{P}\left(|Z| > \left\{k\,SW_c(t^-))\right\}^{1/2}\right),$$
$$\text{for} \quad t \geq 0, \quad \text{and} \quad Z \sim N(0,1).$$

TABLE 4.4: Estimated parameters according to 3.11 and 3.20 formulations. The reader can refer to 3.2 for the utilized formulations of distributions. To enhance the clarity of our presentation, the pair $(\theta_2, k)$ has been aligned in the same column due to their similar placement in their respective equations, as seen in 3.11 and the different formulations of 3.20 discussed in 3.2. Similarly, note that the $a$ parameter means the $a$ exponent in, respectively, the gamma ($a_{gam}$), MLF ($a_{MLF}$) and exp+rate models.

| Version | $\theta_1$ | $\sigma$ | $k$ or $\theta_2$ | $a$ | $\mu_{log}$ | $\sigma_{log}$ | $\theta$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|
| lognorm | 0.011 | 0.165 | - | - | 3.958 | 3.273 | - | - |
| Pareto | 370.112 | 0.165 | - | - | - | - | $3.02 \cdot 10^{-6}$ | 0.403 |
| mlf | 0.01 | 0.166 | 0.017 | 0.501 | - | - | - | - |
| gamma | 0.007 | 0.169 | 0.01 | 0.386 | - | - | - | - |
| exp + rate | 0.007 | 0.172 | 2.037 | $-0.834$ | - | - | - | - |
| GreenLab exp | 0.005 | 0.208 | 0.559 | - | - | - | - | - |
| exp | 0.005 | 0.211 | 0.133 | - | - | - | - | - |

Another noteworthy finding pertains to the Pareto model, specifically the parameter $\eta$. As delineated in Table 4.4, $\eta$ is estimated at 0.403. This value signifies the initial cumulative water consumption $(SW_c(t))$ up to the first observation time, equating to approximately 400 $ml$ over 38 days, or 10.6 $ml$ per day, a value similar to the calculated coefficient of the 'time' variable presented in table 4.1, which represents an increase of 14 $ml$ per day.

TABLE 4.5: Comparison of different distribution choices regarding 3.11 and 3.20 formulations. The table presents the different methods, the calculated log-likelihood value (log_lik_val), RMSE, the total number of parameters, and BIC and AIC criteria. The arrangement of methods is done according to the BIC criterion.

| | Version | log_lik_val | RMSE | # param | BIC | AIC |
|---|---|---|---|---|---|---|
| 1 | lognorm | 20.45 | 0.16 | 4 | **-25.02** | -32.9 |
| 2 | Pareto | 20.39 | 0.16 | 4 | **-24.9** | -32.78 |
| 3 | mlf | 19.85 | 0.17 | 4 | -23.82 | -31.7 |
| 4 | gamma | 19.12 | 0.17 | 4 | -22.36 | -30.24 |
| 5 | exp + rate | 18.15 | 0.17 | 4 | -20.42 | -28.3 |
| 6 | GreenLab exp | 8.14 | 0.21 | 3 | -4.37 | -10.28 |
| 7 | exp | 7.25 | 0.21 | 3 | -2.59 | -8.5 |

Figure 4.3 showcases the computed LIR time-course in relation to the various methodologies discussed in 3.2. Specifically, the pairs *exp-GreenLab exp*, and *gamma-exp + rate* exhibit similar trends.

This similarity is even more visible when the LIR is normalized by its maximal value and displayed with respect to $SW_c$, as shown in Supplementary Material (Appendix 5.1). As the optimization procedure revealed, there is a compensation effect between $\theta_1$ and $LIR$ scaling, thus the interest also of the normalized representation in appendix 3. However, the *Pareto* and *mlf* methodologies demonstrate distinct trends that can be clearly differentiated from the others.

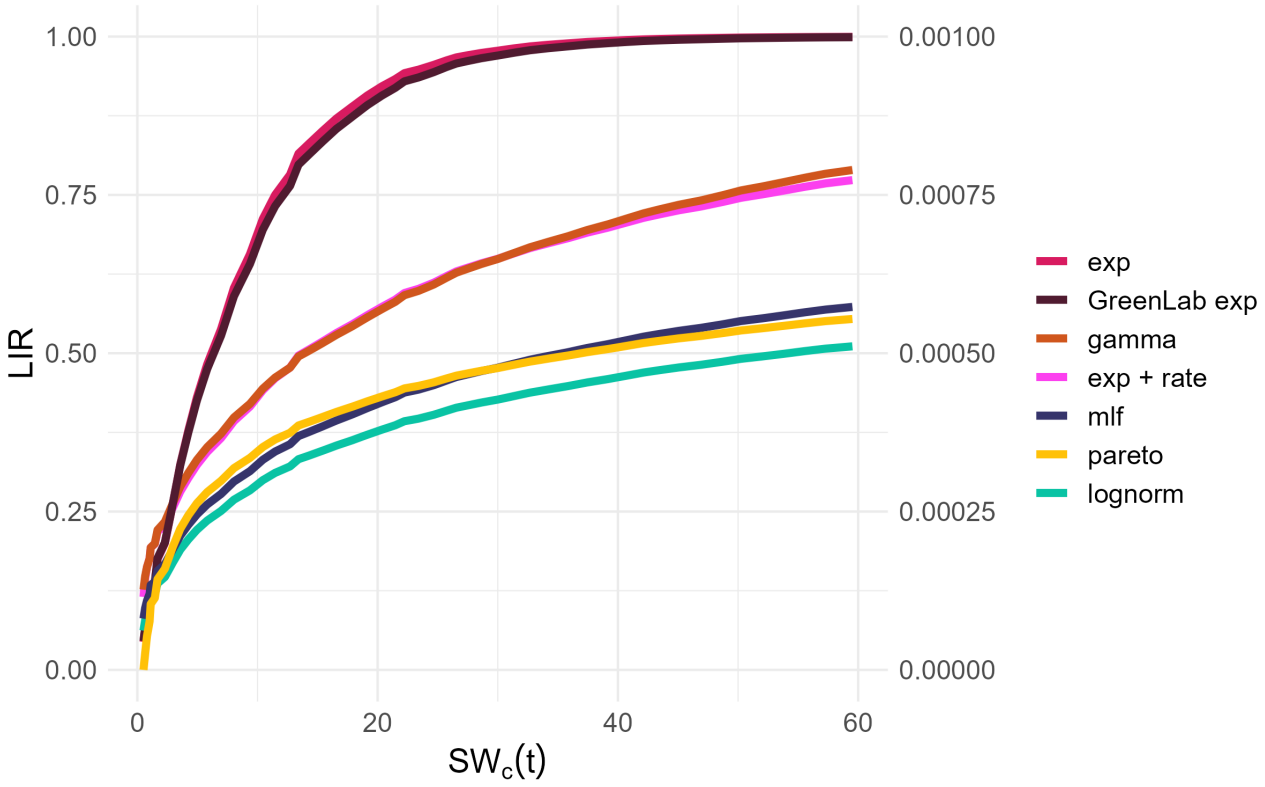


FIGURE 4.3: The estimation of the LIR as a function of the accumulated water usage, as drawn from the suggested models (3.2), is depicted in the provided graph. A second axis was included for the values of pareto distribution. *exp* and *GreenLab exp* are overlapped as are *gamma* and *exp + rate*.

The unique trend of *pareto* methodology is also evident in Figure 4.4, where it manages to track the initial and final trends concurrently during the observation period - a feat unattainable by the other methods, which are only capable of capturing either the beginning or the end trend, but not both simultaneously. Another notable result, concerns the grouping of the best performed methodologies according to the BIC criterion (Figure 4.5), with their position below the remaining ones, depicted at figure 4.3.

## 4.2 Prediction

Table 4.6 encapsulates the outcomes of our predictive analysis. Compared with the results in Table 4.2, our study revealed that the *Pareto*, *lognormal*, and *mlf* models along with the linear models present

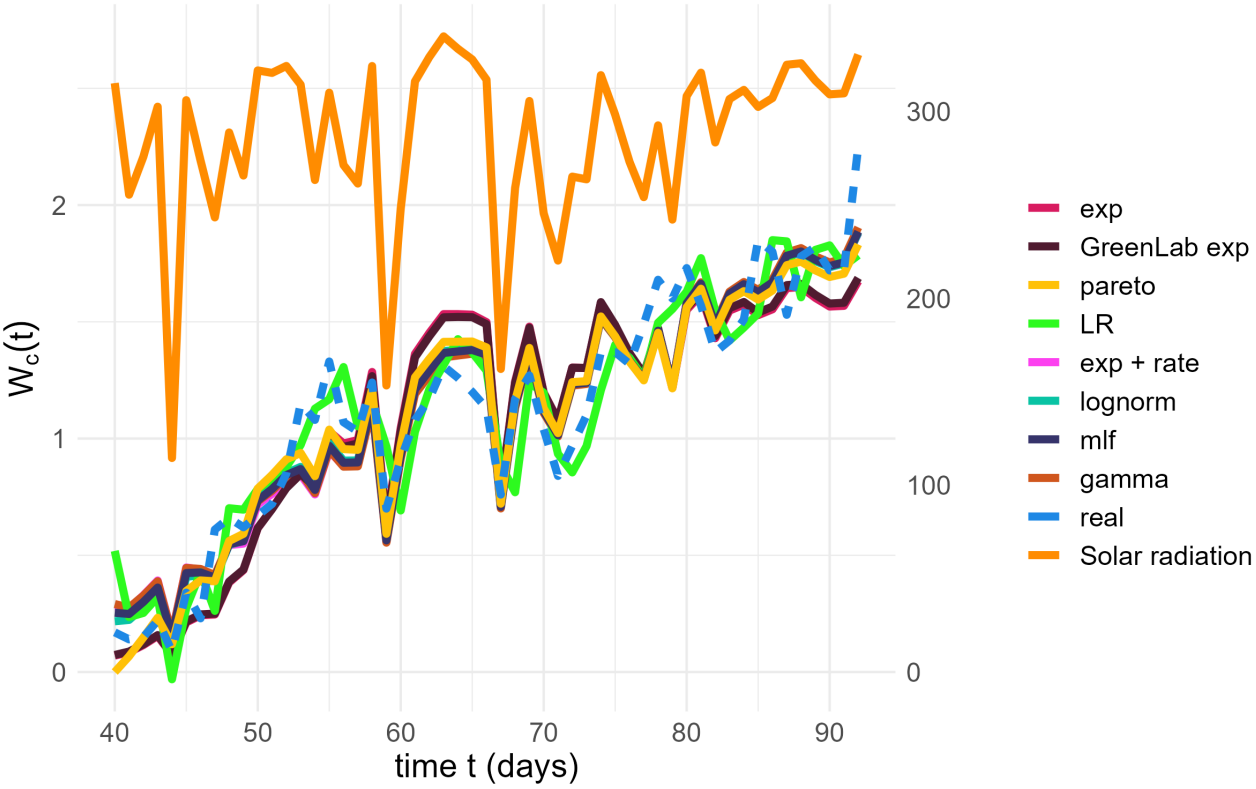


FIGURE 4.4: Final fit of the models (solid lines) on the real data (dashed line). Time (days), represented on the x-axis, runs over the days of observation, with $t = 1$ being the day the seed was planted. The left y-axis represents the Water Consumption at time $t$, in liters. The right y-axis represents values of Avg Solar radiation ($W/m^2$). The evolution of Solar radiation is plotted at the top of the graph, with a dark orange color.

equivalent results under the known Solar Radiation setting, indicating their relative effectiveness within the context of our investigation. However, it is crucial to acknowledge the underperformance of the *exp*, *GreenLab exp* and *exp + rate* models, which implement a methodology similar to the Beer-Lambert law. Compared to other models, these models' inferior performance underscores that specific methodologies might not always be optimal.

TABLE 4.6: Prediction summary among the different suggested methods discussed in formulations 3.11 and 3.20. Solar Radiation is assumed to be known. Methods are compared using the RMSPE. The arrangement is performed under the evaluation of the mentioned criterion.

| | Version | RMSPE |
|---|---|---|
| 1 | Pareto | **0.194** |
| 2 | lognorm | 0.217 |
| 3 | mlf | 0.226 |
| 4 | gamma | 0.234 |
| 5 | GreenLab exp | 0.282 |
| 6 | exp | 0.296 |
| 7 | exp + rate | 0.341 |

# Chapter 5

# Conclusion

During this thesis, we investigated ways to predict Water Consumption of tomato plants (c.v. ecstasis). Our initial goal was to make some heuristic and empirical approaches, currently used by greenhouse producers more rigorous in the mathematical sense. To achieve that, different models were investigated. Among the investigated models, classical linear models (Chapter 1) under a Data-driven scheme and Functional and Structural models (Chapter 3), mainly the GreenLab model, under a Knowledge-driven scope, were investigated. This investigation gave, as a result, the introduction of another family of models, namely the Stochastic Segmentation of input Energy (SSiE), discussed in section 3. This modeling approach combines elements of both approaches. Even though linear models lack biological representation, and the GreenLab model was deemed unidentifiable in our setting (Section 2.3), the SSiE models presented optimistic results (Section 4.2) in prediction, providing also some biological representation for the interception of light (Figure 4.3), and specifically a vague description of Light Interception Ratio (LIR). Unfortunately, our shortage of more informative data, which ideally would include light interception measurements, condemned our work as mainly theoretical. We hope our work will motivate research in the field and applications of the presented methodology in more practical schemes.

## 1 Summary of our main findings

Our current method, employing *Pareto* and *mlk*, yielded comparable predictive outcomes (RMSPE 0.19-0.23). Even lower results were found by the application of the linear models (Table 4.2), even though there is a lack of biological interpretation for this approach, and the assumption of knowledge of the values for a big set of the variables, especially in the model chosen by the backward elimination, could have influenced these results. We believe that even under these limitations, studying the results of these simple models, and especially the ones found in Table 4.1, can give sounder intuition on the phenomenon under observation.

In the context of our problem formulation, which involves one measurement of Water Consumption per day and relies solely on climatic data, this RMSPE translates to an error in the range of 150-250ml per day. This level of accuracy can contribute to the sustainability of agricultural practices by optimizing water usage. Importantly, the *Pareto* and *mlf* models are feasible for application in a scheme of one measurement per day. However, both of them have disadvantages. The Pareto model presents some identifiability issues among the $\mu_0$ and $\theta$ parameters, which warrants further investigation. On the other hand, the *mlf* is computationally heavy, a disadvantage that can be minimal in a scenario with only one measurement and only one day to predict. Despite these challenges, the models remain viable choices for real-world applications. Within this context, when the primary objective is focused on prediction, linear models (Table 4.1) offer a marginal benefit. Even though *exp + rate* and *gamma* models do not present equivalent results as the aforementioned, the LIR estimated by these methods (Figure 4.3), approximately 80%, are similar to the results reported in Wilson et al. (1992) and Ohashi et al. (2022). Measurements at 7 farms showed that in the summer season, the light interception was on average 90%, with values varying between 86% and 96%" in Heuvelink et al. (2004), with reported densities of 2.3 and 3.4 stems per $m^2$, in contrast to our case, where the reported density is 5 stems per $m^2$

Our work can be considered as a methodological proposition for determining the LIR profile with only a subset of the variables routinely measured by professional growers, in a hydroponic setting, variables which are discussed in Chapter 2. Interestingly, the profiles of LIR, Figure 4.3, we obtain are consistent with those reported in the literature ((Duursma et al., 2012), Ohashi et al. (2022)). Selecting the model with the best predictive performances seems a reasonable strategy. Nonetheless, this approach warrants further empirical validation. Future research could focus on quantifying the diffusion of light in relation to distinct plant attributes and may include virtual experiments (as in (Duursma et al., 2012)).

## 2 Limitations of the work

Our work presents important limitations that must be acknowledged. First, our modeling approach relies on strong physiological simplifications, e.g., neglecting soil evaporation and respiration of existing organs, constant radiation use efficiency, proportionality between water consumption and biomass production, constant SLA, proportionality between light intercepted and photosynthesis (a more refined model here would have been to consider Farquhar's photosynthesis model, for instance Farquhar et al. (1980)). All these simplifications were required with respect to our objectives and our context of using only routinely recorded variables. They can, however, be considered applicable when

describing the average growth of plants in standard conditions, and most of them are also laid in other models (Ma et al. (2022),Winn et al. (2023)).

An additional underlying assumption that deserves to be highlighted is that the $g$ function (discussed in paragraph 3.1) is time-independent. In reality, $g$ aggregates the effects of blade spatial arrangement, which determines the probability of intercepting a radiation ray, the senescence of the leaves, and the fraction of biomass allocated to the blades. This fraction decreases with time, especially due to the progressive appearance of fruits, whose demand competes with that of blades, a phenomenon that our SSiE models do not account for. However, in our case, because the time of observation is at a very later stage than the initial planting, this fraction is, in fact, nearly constant, taking values in the range (0.21-0.24), as simulated using GreenLab (Zhang et al., 2009). This explains why the models *exp* and *GreenLab exp* behave similarly.

Lastly, we must acknowledge the limitations of our data, which prevent us from drawing strong conclusions from our results. Measuring and estimating the mean value of water consumption among three plants could potentially introduce some errors because of the variance within them. Nevertheless, we believe that our work can be considered a proof-of-concept for our proposed methodology and that the SSiE model appears promising for modeling Water Consumption.

## 3 Perspectives

In light of this, future research could aim to further apply and investigate the utility of the SSiE models in predicting such quantities. The choice of distribution might be crop-dependent, an idea that could be researched in the future. We hope these initial findings can be validated with more extensive and informative data and deepen our understanding of crop Water Consumption patterns.

Our current formulation is particularly adapted for Bayesian methods, which will allow for an easy way to quantify uncertainty and use the Bayesian predictive distribution for forecasting purposes. An online Bayesian method with sequential Monte-Carlo may be particularly relevant, and MCMC methods could also be applied for more efficient estimation, as in (Logothetis et al., 2022). The comparison of MCMC with sequential Monte-Carlo for MLE was done in Trevezas et al. (2014).

We anticipate that this method could have repeated applications within the same crop type and its application to other crops and settings. These possibilities present exciting avenues for future work and could have far-reaching impacts on sustainable agriculture and food security.

# Bibliography

H. Bateman. *Higher transcendental functions [volumes i-iii]*, volume 1. McGRAW-HILL book company, 1953.

L. Boltzmann. *Theoretical physics and philosophical problems: Selected writings*, volume 5. Springer Science & Business Media, 2012.

K. J. Boote, J. W. Jones, J. W. White, S. Asseng, and J. I. Lizaso. Putting mechanisms into crop production models. *Plant, cell & environment*, 36(9):1658–1672, 2013.

H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

A. L. Buck. New equations for computing vapor pressure and enhancement factor. *Journal of Applied Meteorology and Climatology*, 20(12):1527–1532, 1981.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

B. W. Carroll and D. A. Ostlie. An introduction to modern astrophysics and cosmology. *An introduction to modern astrophysics and cosmology/BW Carroll and DA Ostlie. 2nd edition. San Francisco: Pearson*, 2006.

G. Casasanta and R. Garra. Towards a generalized beer-lambert law. *Fractal and Fractional*, 2(1):8, 2018.

F. Castellvi, P. Perez, J. Villar, and J. Rosell. Analysis of methods for estimating vapor pressure deficits and relative humidity. *Agricultural and Forest Meteorology*, 82(1-4):29–45, 1996.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

Y. Chen, S. Trevezas, and P.-H. Cournède. Iterative convolution particle filtering for nonlinear parameter estimation and data assimilation with application to crop yield prediction. In *2013 Proceedings of the Conference on Control and its Applications*, pages 67–74. SIAM, 2013.

P. De Reffye, B. Hu, M. Kang, V. Letort, and M. Jaeger. Two decades of research with the greenlab model in agronomy. *Annals of Botany*, 127(3):281–295, 2021.

Q. Dong, G. Louarn, Y. Wang, J.-F. Barczi, and P. De Reffye. Does the structure–function model greenlab deal with crop phenotypic plasticity induced by plant spacing? a case study on tomato. *Annals of botany*, 101(8):1195–1206, 2008.

R. A. Duursma, D. S. Falster, F. Valladares, F. J. Sterck, R. W. Pearcy, C. H. Lusk, K. M. Sendall, M. Nordenstahl, N. C. Houter, B. J. Atwell, et al. Light interception efficiency explained by two simple variables: a test using a diversity of small-to medium-sized woody plants. *New Phytologist*, 193(2):397–408, 2012.

N. K. Fageria. *The role of plant roots in crop production*. CRC Press, 2012.

J. J. Faraway. *Practical regression and ANOVA using R.*, volume 168. University of Bath Bath, 2002.

G. D. Farquhar, S. v. von Caemmerer, and J. A. Berry. A biochemical model of photosynthetic co 2 assimilation in leaves of c 3 species. *planta*, 149:78–90, 1980.

E. Fermi. Thermodynamics dover publications. *New York*, 1956.

K. V. Flannery. The origins of agriculture. *Annual review of Anthropology*, 2(1):271–310, 1973.

Food and Agriculture Organization of the United Nations. *Crop evapotranspiration - Guidelines for computing crop water requirements*. Food and Agriculture Organization of the United Nations, 1998. URL `https://www.fao.org/3/X0490E/x0490e00.htm#Contents`. Accessed: 2023-07-12.

C. A. Gueymard. The sun's total and spectral irradiance for solar energy applications and solar radiation models. *Solar energy*, 76(4):423–453, 2004.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

J. L. Hatfield and J. H. Prueger. Temperature extremes: Effect on plant growth and development. *Weather and climate extremes*, 10:4–10, 2015.

H. J. Haubold, A. M. Mathai, R. K. Saxena, et al. Mittag-leffler functions and their applications. *Journal of applied mathematics*, 2011, 2011.

A. M. Hetherington and F. I. Woodward. The role of stomata in sensing and driving environmental change. *Nature*, 424(6951):901–908, 2003.

E. Heuvelink, M. Bakker, A. Elings, R. Kaarsemaker, and L. Marcelis. Effect of leaf area on tomato yield. In *International Conference on Sustainable Greenhouse Systems-Greensys2004 691*, pages 43–50, 2004.

T. Howell and J. Musick. Relationship of dry matter production of field crops to water consumption. In *Les besoins en eau des cultures, conférence internationale, Paris, Versailles, 11-14 septembre 1984*, pages 247–269, 1985.

T. Howell, K. Davis, R. McCormick, H. Yamada, V. T. Walhood, and D. Meek. Water use efficiency of narrow row cotton. *Irrigation Science*, 5:195–214, 1984.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

H. Jones. Energy balance and evaporation. *Plants and Microclimate. A quantitative approach to environmental plant physiology, 2nd edition*, pages 106–130, 1992.

J. M. Juran and J. A. De Feo. *Juran's quality handbook: the complete guide to performance excellence*. McGraw-Hill Education, 2010.

S. A. Kalogirou. Solar thermal collectors and applications. *Progress in energy and combustion science*, 30 (3):231–295, 2004.

G. Kopp and J. L. Lean. A new, lower value of total solar irradiance: Evidence and climate significance. *Geophysical Research Letters*, 38(1), 2011.

J. Lanoue, E. D. Leonardos, X. Ma, and B. Grodzinski. The effect of spectral quality on daily patterns of gas exchange, biomass gain, and water-use-efficiency in tomatoes and lisianthus: An assessment of whole plant measurements. *Frontiers in plant science*, 8:1076, 2017.

W. Larcher. *Physiological plant ecology: ecophysiology and stress physiology of functional groups*. Springer Science & Business Media, 2003.

V. Letort, P.-H. Cournède, and P. De Reffye. Impact of topology on plant functioning: a theoretical analysis based on the greenlab model equations. In *2009 Third International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications*, pages 341–348. IEEE, 2009.

J. Levitt. Response of plants to environmental stresses: chilling, freezing, and high temperature stresses. *Physiological ecology: a series of monographs, texts, and treatises*, 1:23–64, 1980.

D. B. Lobell, W. Schlenker, and J. Costa-Roberts. Climate trends and global crop production since 1980. *Science*, 333(6042):616–620, 2011.

D. Logothetis, S. Malefaki, S. Trevezas, and P.-H. Cournède. Bayesian estimation for the greenlab plant growth model with deterministic organogenesis. *Journal of Agricultural, Biological and Environmental Statistics*, 27(1):63–87, 2022.

J. Lu, G. Sun, S. G. McNulty, and D. M. Amatya. A comparison of six potential evapotranspiration methods for regional use in the southeastern united states 1. *JAWRA Journal of the American Water Resources Association*, 41(3):621–633, 2005.

N. Lu, T. Nukaya, T. Kamimura, D. Zhang, I. Kurimoto, M. Takagaki, T. Maruo, T. Kozai, and W. Yamori. Control of vapor pressure deficit (vpd) in greenhouse enhanced tomato growth and productivity during the winter season. *Scientia Horticulturae*, 197:17–23, 2015.

C. Ma, M. Liu, F. Ding, C. Li, Y. Cui, W. Chen, and Y. Wang. Wheat growth monitoring and yield estimation based on remote sensing data assimilation into the safy crop growth model. *Scientific Reports*, 12(1):5473, 2022.

J. L. Monteith. Climate and the efficiency of crop production in britain. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 281(980):277–294, 1977.

Y. Ohashi, M. Murai, Y. Ishigami, and E. Goto. Light-intercepting characteristics and growth of tomatoes cultivated in a greenhouse using a movable bench system. *Horticulturae*, 8(1):60, 2022.

Oxford-University-Press. Oxford english dictionary, 1989.

R. N. Pillai. On mittag-leffler functions and related distributions. *Annals of the Institute of statistical Mathematics*, 42:157–161, 1990.

C. R. Pivetta, I. F. Tazzo, G. F. Maass, N. A. Streck, and A. B. Heldwein. Leaf emergence and expansion in three tomato (lycopersicon esculentum mill.) genotypes. *Ciência Rural*, 37:1274–1280, 2007.

J. R. Porter and M. Gawith. Temperatures and the growth and development of wheat: a review. *European journal of agronomy*, 10(1):23–36, 1999.

H. M. Resh. *Hydroponic food production: a definitive guidebook for the advanced home gardener and the commercial hydroponic grower*. CRC press, 2022.

D. Schmidt, D. T. Zamban, D. Prochnow, B. O. Caron, V. Q. Souza, G. M. Paula, C. Cocco, et al. Phenological characterization, phyllochron and thermal requirement of italian tomato in two cropping seasons. *Horticultura Brasileira*, 35(1):89–96, 2017.

J. Schönherr. Characterization of aqueous pores in plant cuticles and permeation of ionic solutes. *Journal of Experimental Botany*, 57(11):2471–2491, 2006.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

R. R. Shamshiri, J. W. Jones, K. R. Thorp, D. Ahmad, H. C. Man, and S. Taheri. Review of optimum temperature, humidity, and vapour pressure deficit for microclimate evaluation and control in greenhouse cultivation of tomato: a review. *International agrophysics*, 32(2):287–302, 2018.

R. L. Snyder, D. Spano, C. Cesaraccio, and P. Duce. Determining degree-day thresholds from field observations. *International Journal of Biometeorology*, 42:177–182, 1999.

R. B. Stull. *Practical meteorology: an algebra-based survey of atmospheric science*. University of British Columbia, 2015.

L. Taiz, E. Zeiger, I. M. Møller, A. Murphy, et al. *Plant physiology and development.* Sinauer Associates Incorporated, 6 edition, 2015.

K. E. Trenberth, J. T. Fasullo, and T. G. Shepherd. Attribution of climate extreme events. *Nature Climate Change*, 5(8):725–730, 2015.

S. Trevezas, S. Malefaki, and P.-H. Cournède. Parameter estimation via stochastic variants of the ecm algorithm with applications to plant growth modeling. *Computational Statistics & Data Analysis*, 78: 82–99, 2014.

M. Ünlü, R. Kanber, D. L. Koç, S. Tekin, and B. Kapur. Effects of deficit irrigation on the yield and yield components of drip irrigated cotton in a mediterranean environment. *Agricultural Water Management*, 98(4):597–605, 2011.

D. Van der Zande, J. Stuckens, W. W. Verstraeten, B. Muys, and P. Coppin. Assessment of light environment variability in broadleaved forest canopies using terrestrial laser scanning. *Remote Sensing*, 2(6):1564–1574, 2010.

J. M. Wallace and P. V. Hobbs. *Atmospheric science: an introductory survey*, volume 92. Elsevier, 2006.

J. W. Wilson, D. Hand, and M. Hannah. Light interception and photosynthetic efficiency in some glasshouse crops. *Journal of Experimental Botany*, 43(3):363–373, 1992.

C. A. Winn, S. Archontoulis, and J. Edwards. Calibration of a crop growth model in apsim for 15 publicly available corn hybrids in north america. *Crop Science*, 63(2):511–534, 2023.

H.-P. Yan, M. Z. Kang, P. De Reffye, and M. Dingkuhn. A dynamic, architectural plant model simulating resource-dependent growth. *Annals of botany*, 93(5):591–602, 2004.

B. Zhang, M. Kang, V. Letort, X. Wang, and P. De Reffye. Comparison between empirical or functional sinks of organs-application on tomato plant. In *2009 Third International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications*, pages 191–197. IEEE, 2009.

# Appendix A: R code

Simulate a cycle function for the simulation of the tomato automaton on greenlab model

```
# p.cyc: prob of a cycle to be realized
# p.fr: probability of a flower to become a fruit
# fl_1st: 1st cycle that a flower blossoms
Sim.cyc_fun <- function(N, p.cyc = 1, p.fr = 1, fl_1st = 9, organs = c("b","
    p","e","f")){
  # standard tomato cycle (b,p,e)
  s.cyc <- c(1,1,1,0); s.cyc
  # Number maximum Cycles of development
  N <- N
  # Number of realized cycles
    N.cyc <- rbinom(1,N,p.cyc); N.cyc


  # simulate cycles
    Sim.cyc <- matrix(s.cyc, ncol = 4, nrow = N.cyc, byrow = T,
                      dimnames = list(cycle = paste0("c_",1:N.cyc),
                                      comp = organs)); Sim.cyc
  # cycle of 1st flower
    fl_1st <- 9
    # every third cycle it flowers
    suppressWarnings(a <- fl_1st:N.cyc*c(T,F,F)); a <- a[a!=0]; a
    # actual fruits developed under p.fr
    Sim.cyc[a,4] <- rbinom(length(a),5,p.fr); Sim.cyc
    Sim.cyc
}
```

GreenLab function

```
library(vctrs) # rep_vec_each function
# Simulate a cycle function
source("./source/simulate a cycle function.R")
# Parameters

par = list( phi = 2,             # Phyllochron
            k = 0.8,              # Beer-Lambert coef of light extinction
            bt = 12,               # base temperature
            GDD=10.16,            # mean GDD per day
            To = 30               # maximum expansion time
)




# Calculate Thermal Time from Calendar Time

thermal.time = function(t,  par = par) {
  par$GDD * t
}

# Cycle of development
# t: time in days
# temp: data of temp at day(s) t
# p
CD <- function(t, par = par){
  one_CD <- par$phi * par$GDD # on cycle of development
  floor(thermal.time(t, par)/one_CD)
}




# Beer - Lambert Law
# variables
```

```
# qleaf: vector with biomass of leaf

beer.lambert = function(qleaf, ET, theta) {
  q = unname( ET * theta["Sp"] / ( theta["r1"]) * (1 - exp( - par$k* sum(
     qleaf)/( theta["Sp"] * theta["e"]))) )
  q
}

# Demand of organ o (b:blade, p: petiole, e: internode, f:fruit)
# variables
# j: time since initiation of organ o
# B: parameter of Beta(a-1,b-1) B=a/(a+b), a+b=5
# P: sink strength of organ o
# D: Total plant demand at time t
# Beta law

g <- function(j, A, B, To){
  a <- A
  b <- B
  (j-1/2)^(a-1)*(To - j + 1/2)^(b-1)
}

# sink variation function

f <- function(j, A, B, To){

  (g(j, A, B,To)/sum(g(1:To, A, B,To)))*I(j<=To)

}



# Greenlab
# t: time in days
# theta: parameter  under estimation
# ET: evapotranspiration (or if adjusted accordingly other environmental
   parameter (i.e. Solar radiation)
greenlab <-  function(t, theta, par, ET){
```

```
organs <- c("b","p","e","f") # organ o (b:blade, p: petiole, e: internode,
    f: fruit)
# simulate until CD(t)
Sim.cyc <- as.data.frame(Sim.cyc_fun(CD(t, par), organs = organs)); Sim.
   cyc
# list with indexes
ind <- list(N = nrow(Sim.cyc)+1) # +1 for the "seed" cycle
for (o in organs) {
  ind[paste0("n_",o)] <- list(sum(Sim.cyc[,o]))
}

## Initialize ##
# Note: extra variables only for better Latex presentation
# otherwise they are defined immediately in x
# total biomass
q <- unname(rep(theta["q0"], times=ind$N))
# total demand
D <- rep(0, times=ind$N)
# demand per compartment
d <- list(b = matrix(0, nrow=ind$N, ncol=ind$n_b,
                     dimnames = list(cycle = paste0("cyc",
                                                    1:ind$N),
                                     compartment = paste0("comp",
                                                          1:ind$n_b))),
          p = matrix(0, nrow=ind$N, ncol=ind$n_p),
          dimnames = list(cycle = paste0("cyc", 1:ind$N),
                          compartment = paste0("comp",
                                               1:ind$n_p)),
          e = matrix(0, nrow=ind$N, ncol=ind$n_e),
          dimnames = list(cycle = paste0("cyc", 1:ind$N),
                          compartment = paste0("comp",
                                               1:ind$n_e)),
          f= matrix(0, nrow=ind$N, ncol=ind$n_f),
          dimnames = list(cycle = paste0("cyc", 1:ind$N),
                          compartment = paste0("comp",
                                               1:ind$n_f)))
# dry matter allocation
```

```
dq_o <- list(b = matrix(0, nrow=ind$N, ncol=ind$n_b,
                        dimnames = list(cycle = paste0("cyc",
                                                       1:ind$N),
                                        compartment = paste0("comp",
                                                             1:ind$n_b)
                        )),
             p = matrix(0, nrow=ind$N, ncol=ind$n_p),
             dimnames = list(cycle = paste0("cyc", 1:ind$N),
                             compartment = paste0("comp", 1:ind$n_p)),
             e = matrix(0, nrow=ind$N, ncol=ind$n_e),
             dimnames = list(cycle = paste0("cyc", 1:ind$N),
                             compartment = paste0("comp", 1:ind$n_e)),
             f= matrix(0, nrow=ind$N, ncol=ind$n_f),
             dimnames = list(cycle = paste0("cyc", 1:ind$N),
                             compartment = paste0("comp", 1:ind$n_f)
             ))
# biomass per compartment
q_o <- list(b = matrix(0, nrow=ind$N, ncol=ind$n_b,
                       dimnames = list(cycle = paste0("cyc", 1:ind$N),
                                       compartment = paste0("comp",
                                                            1:ind$n_b))),
            p = matrix(0, nrow=ind$N, ncol=ind$n_p),
            dimnames = list(cycle = paste0("cyc", 1:ind$N),
                            compartment = paste0("comp", 1:ind$n_p)),
            e = matrix(0, nrow=ind$N, ncol=ind$n_e),
            dimnames = list(cycle = paste0("cyc", 1:ind$N),
                            compartment = paste0("comp", 1:ind$n_e)),
            f= matrix(0, nrow=ind$N, ncol=ind$n_f),
            dimnames = list(cycle = paste0("cyc", 1:ind$N),
                            compartment = paste0("comp", 1:ind$n_f)))

# total list for export
x = list(q = q, # repeat initial biomass
         D = D,
         d = d,
         dq_o = dq_o,
         q_o = q_o,
         t = 1:t,CD = 1:CD(t,par), ind = ind)
```

```
# start simulating from the first cycle
for (n in 1:(ind$N-1)){

  # current cycle
  age <- n:1; age
  curr.cyc <- cbind(Sim.cyc[1:n,],age); curr.cyc # keep info till current
     cycle
  Demand <- 0
  for (o in organs) {
    # calculate sink function (here To is assumed equal everywhere, can be
       replace by T[o] for general)
    curr.cyc[,"Fo"] <- f(age,theta[paste0("A",o)],theta[paste0("B",o)],
       par$To); curr.cyc

    curr.n_o <- sum(curr.cyc[,o]); curr.n_o # current number of organ

    zero.demand <- rep(0,ind[[paste0("n_",o)]]-curr.n_o); zero.demand #
       demand of organs not elapsed yet

    # calculate demand of each compartment
    if (curr.n_o == 0){
      x$d[[o]][n,] <- 0
    } else {
      Fo <- vec_rep_each(curr.cyc[curr.cyc[,o]>0,c("Fo")], times = curr.
         cyc[curr.cyc[,o]>0,o]); Fo[is.nan(Fo)] <- 0; Fo
      dem <- theta[paste0("P",o)]*Fo; dem
      x$d[[o]][n,] <- c(dem,zero.demand)
    }
    Demand <- Demand + sum(x$d[[o]][n,]) # Total demand
  }
  x$D[n] <- Demand

  if (!is.na(Demand) && Demand > 0){
    for (o in organs) {
      # Dry matter allocation for each compartment
      x$dq_o[[o]][(n+1),] <- x$d[[o]][n,]/Demand*x$q[n]
      # new biomass for each compartment
```

```
        x$q_o[[o]][(n+1),] <- x$q_o[[o]][n,] + x$dq_o[[o]][(n+1),]
      }
    }


    # Beer lambert for new biomass
    x$q[n+1] = beer.lambert(qleaf=x$q_o[["b"]][n,], ET = ET[n+1], theta)
  }

  x
}
```

# Appendix B: Main Dataset

TABLE 5.1: Summary of all the measured quantities. All quantities refer to a statistic measured on a day span. In the first column, the name of the quantity can be found along with the units. Columns 2 to 6 present statistics of the quantities along the span of the study (54 days).

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| time ($t$) | 54 | 31.412 | 15.292 | 6 | 57 |
| Max_Solar ($W/m^2$) | 54 | 1,025.431 | 132.998 | 571 | 1,329 |
| Avg_ET ($mm/m^2$) | 54 | 4.608 | 0.990 | 2.350 | 7.490 |
| Max_ET ($mm/m^2$) | 54 | 8.216 | 1.591 | 4 | 12 |
| Min_ET ($mm/m^2$) | 54 | 1.098 | 0.781 | 0 | 4 |
| Avg_Air_pressure ($hPa$) | 54 | 1,013.817 | 3.477 | 1,007.120 | 1,022.720 |
| GDDs ($^oCd$) | 54 | 10.156 | 2.668 | 5.030 | 16.630 |
| Avg_Temp ($^oC$) | 54 | 22.156 | 2.675 | 17.060 | 28.630 |
| Max_Temp ($^oC$) | 54 | 29.292 | 3.145 | 23.100 | 37.800 |
| Min_Temp ($^oC$) | 54 | 15.927 | 3.235 | 9.400 | 22.000 |
| Avg_Hum (%) | 54 | 0.843 | 0.103 | 0.618 | 0.976 |
| Max_Hum (%) | 54 | 0.986 | 0.030 | 0.840 | 1.000 |
| Min_Hum (%) | 54 | 0.597 | 0.156 | 0.230 | 0.860 |
| Avg_VPD ($kPa$) | 54 | 0.547 | 0.385 | 0.090 | 1.370 |
| Max_VPD ($kPa$) | 54 | 1.671 | 0.900 | 0.540 | 3.960 |
| Min_VPD ($kPa$) | 54 | 0.022 | 0.046 | 0.000 | 0.260 |
| Water_Consumption ($L$) | 54 | 1.141 | 0.503 | 0.090 | 2.250 |
| Cut_leaves (Indicator) | 54 | 0.176 | 0.385 | 0 | 1 |
| past_wc ($L$) | 53 | 1.096 | 0.497 | 0.090 | 1.850 |
| past_wc_2days ($L$) | 52 | 1.064 | 0.499 | 0.090 | 1.850 |
| Solar Radiation (standardize) | 54 | 0.865 | 0.136 | 0.360 | 1.000 |

# Supplementary figure: normalized LIR w.r.t. cumulated water uptake

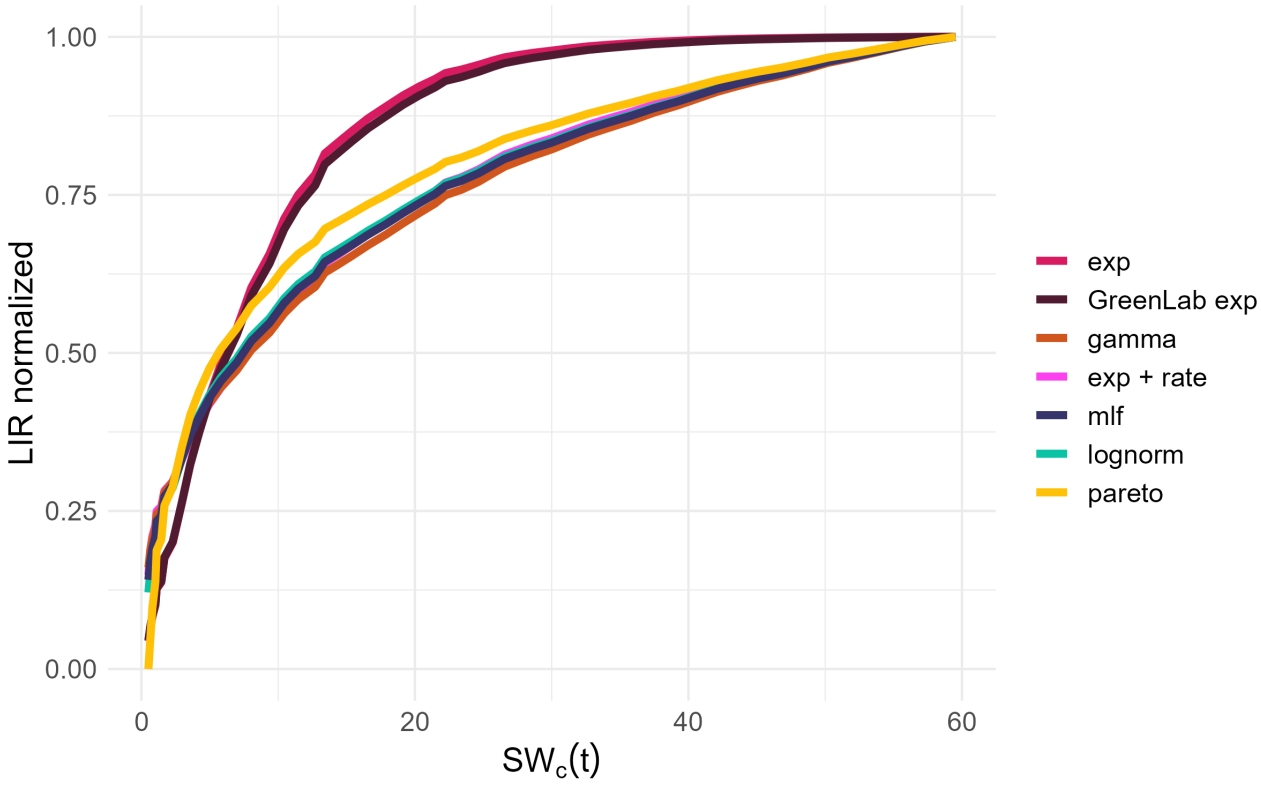


FIGURE 5.1: The estimation of the normalized LIR as a function of the accumulated water usage, as drawn from the suggested models (3.2).

# Index