UPPSALA
UNIVERSITET

Evaluating Resilience to Heat Stress among Dairy Cows in Sweden

A Gigacow Project at SLU

Joakim Svensson & Axel Englund

Project in Computational Science: Report

February 2024

PROJECT REPORT

DEPARTMENT OF INFORMATION TECHNOLOGY

Abstract

This study examines how Swedish weather conditions, especially heat, impact dairy cows' milk production on Swedish farms. Using weather data from Sveriges meteorologiska och hydrologiska institut (SMHI) and dairy data from the extensive Gigacow project from Sveriges lantbruksuniversitet (SLU), a range of mathematical and machine learning methods were employed. Techniques like normalization based on Wood's Lactation Curve and Random Forest were used for in-depth data analysis and predictive modelling, while Bayesian regression provided a statistical framework for understanding parameter distributions and the uncertainties involved. These methods not only demonstrated significant temperature impacts on the milk yield but also underscored the need for farm-specific strategies in addressing heat stress. They clearly showed a pattern where milk production starts to decline once temperatures exceed 15°C. This study aimed to provide insights and tools for further studies that eventually will provide recommendations for farm practices and a decision basis for breeding selection.

Contents

1 Introduction **3**

2 Previous Work **3**

3 Theory **3**

3.1 The Influence of Weather on Dairy Cow Productivity 3

3.2 Analysis of Milk Production Patterns in Dairy Cow Herds 6

3.3 Random Forest 6

3.4 Bayesian Linear Regression 8

4 Implementation **11**

4.1 Datasets 11

4.2 Basic Statistics 15

4.3 Normalize the Daily Yield Data using Wood's Lactation Curve 15

4.4 Implementation of Random Forest 15

4.5 Implementation of Bayesian Linear Regression 16

5 Results **18**

5.1 Basic Statistics 18

5.2 Results for the Implementation of Random Forest 19

5.3 Results for the Implementation of Bayesian Regression 22

6 Discussion **27**

6.1 Summary of Results 27

6.2 Discussion of Results 28

7 Conclusions & Further Work **30**

Acknowledgements **31**

References **32**

Appendix **34**

A Theory **34**

B Method **35**

1 Introduction

This study explores how weather affects dairy production and the health of dairy cows in Sweden. It is driven by recent progress in collecting data from dairy farms across the country, namely, the Gigacow project at Sveriges lantbruksuniversitet (SLU), along with detailed weather data from Sveriges meteorologiska och hydrologiska institut (SMHI). This rich dataset allows for a more in-depth analysis than before.

We focus on the specific challenges of Sweden's temperate climate, a topic less extensively covered in the existing body of research. Most studies have looked at more tropical climates, leaving a gap in our understanding of milder climates' effects on dairy farming. Gigacow aims to fill this gap and enhance our understanding of how weather and milk production interact under these conditions.

Considering the ongoing issue of climate change, we hope this research can provide some useful insights. It aims to add to the discussion on sustainable agriculture and animal welfare and might have practical outcomes for dairy farming in similar climates. By identifying patterns linking heat, cow health, and dairy production, we hope further work can be done based on the different approaches and models we used in this project, to offer suggestions for farming practices and breeding choices to better cope with climate-related changes.

In terms of methods, we use a mix of statistical and machine-learning techniques to filter through the data, looking for patterns and relationships that could lead to helpful discoveries.

We hope the results of this study can offer recommendations to improve the health and well-being of dairy cows, which could in turn help make dairy farming more economically sustainable in the face of climate change.

2 Previous Work

Leading into this inquiry is the study ”Evaluating Locally Measured Weather Variables and Their Effect on Dairy Cow Productivity” from 2021 [1], which examined the precision of local weather measurements in assessing their utility for dairy farm management. This work highlighted the potential for simple, localized weather data to inform farm practices, suggesting that even non-complex data collection methods could have substantial benefits.

Following in chronological order and building upon these findings is the 2022 study ”The Impact of Weather Factors on the Productivity of Dairy Cows” [2]. This subsequent research broadened the scope by integrating national weather data with detailed milking statistics, uncovering a more complex picture of the temperature-milk yield relationship. The analysis identified a notable dependency of milk production on temperature variations, which could be partially masked by the cows' lactation cycles.

This study seeks to expand on these prior analyses by incorporating a wider array of aspects of the weather and a more intricate data analysis methodology. The aim is to refine the conclusions drawn by previous research, with a particular focus on making it more evident how heat stress impacts cows by examining their milk production. By doing so, the research hopefully will enhance the predictive models used by farmers and suggest new avenues for achieving sustainable and economically viable dairy farming operations, while also highlighting the importance of managing heat stress in dairy herds.

3 Theory

This section explores how varying weather conditions, particularly temperature fluctuations and heat stress, affect the productivity and well-being of dairy cows. Insights from recent research highlight the complexity of these interactions and underline the need for adaptive strategies in dairy farming. Followed by the theoretical foundation used by the statistical methods applied to this problem.

3.1 The Influence of Weather on Dairy Cow Productivity

Recent studies indicate that dairy cows in Sweden experience a decrease in milk production at lower temperatures than previously assumed. Specifically, production begins to decline at around 15°C and is significantly affected when the maximum daily temperature reaches approximately 20°C [3]. This sensitivity is partly due to the physiological demands of milk production, likening dairy cows to marathon runners who require optimal conditions for peak performance. Ideal temperatures for dairy cows range between 5°C and 10°C, with a comfort zone extending from -15°C to +25°C. However, these ranges do not account for individual variances among cows or the impact of other environmental factors like humidity, wind, and sunlight [4].

Heat Stress & Its Consequences

Heat stress in dairy cows leads to a range of acute and long-term effects. Acutely, cows exhibit lethargy, loss of appetite, reduced movement, and signs of frustration, akin to how humans respond to discomfort from heat. Over time, heat stress impacts reproductive capabilities, making it difficult for cows to conceive [5]. This challenge was particularly evident during the summer of 2018, which was an abnormally hot summer in Sweden, where fertility rates dropped noticeably [6]. Furthermore, unborn calves exposed to heat stress in utero are prone to lower milk production later in life, shorter lifespans, and passing these effects onto their offspring [7].

Day Temperature- & Night Temperature Categories

Daytime temperatures are categorized based on their physiological impact on dairy cows, crucial for their well-being and productivity. The categories relevant to heat stress are [4]:

- **Cold** (Below 0°C): Increased energy for warmth.
- **Mild Cold** (0°C to 5°C): Generally comfortable.
- **Ideal** (5°C to 15°C): Optimal for health and production.
- **Comfort** (15°C to 20°C): Comfortable, minimal stress.
- **Caution** (20°C to 25°C): Onset of heat stress.
- **Alert** (25°C to 30°C): Increased heat stress risk.
- **Danger** (Above 30°C): Severe heat stress.

Based on the daytime categories, nighttime temperatures are adjusted to reflect cooler conditions, since it is colder during the nights, important in Sweden's climate where tropical conditions are rare. These are:

- **Cold** (Below 0°C): Increased energy for warmth.
- **Mild** (0°C to 5°C): Comfortable for cows.
- **Ideal** (5°C to 15°C): Perfect for cows.
- **Warm** (15°C to 20°C): Increased heat stress risk.
- **Tropical** (Above 20°C): Severe heat stress.

These categorizations are important for managing the impact of temperature fluctuations on dairy cows, especially during the summer months [8, 9, 4].

Economic Implications & Adaptive Strategies

The economic impact of heat stress on dairy farming is significant. For instance, the summer of 2018, saw national losses in Sweden amounting to nine million kronor due to reduced milk production, not accounting for other heat-related losses like decreased fertility and udder health. On a farm level, the strategies to lighten heat stress vary widely, with some farms doing better than others. Which strategies are the most efficient remains unclear but this will likely vary depending on specific farm conditions [3]. Common practices include providing shade, using fans, and wetting cows to alleviate heat stress. An alternative approach involves night grazing, allowing cows to feed outdoors during cooler temperatures [10].

Future Strategies, Outlook & Research

The strategies adopted by farms vary widely, with a primary focus on maintaining milk production and udder health, rather than preventing the consequences of reduced fertility. However, the effects of decreased fertility during summer are evident, as seen in unevenly distributed calvings throughout the year, which disrupt the flow of animals on the farms. Failed inseminations and unobserved estrus cycles are direct consequences of heat stress on the reproductive capabilities of cows [6].

The impacts on pregnant cows are less obvious and often overlooked. Research from the USA indicates that cows exposed to heat stress during pregnancy experience reduced milk production in subsequent lactations. Additionally, the calves born to these cows tend to have lower milk production, reduced fertility, shorter lifespans, and a likelihood of passing on the trait of lower milk production to their offspring [7].

Looking ahead, understanding individual differences in heat tolerance among cows could inform future breeding programs aimed at enhancing resilience to heat stress. This approach may be crucial for maintaining sustainable dairy production in the face of increasingly warmer climates [11].

Lactation Cycles

The lactation cycle in cows is the period from when a cow gives birth to when it stops producing milk, usually lasting about 305 to 340 days. It begins with a quick rise in milk production right after the cow has a calf. This peaks about six weeks later but then gradually decreases until the cow stops making milk. This cycle is energy-intensive, demanding substantial food intake and nutritional support. However, cows often face limitations in food intake, leading to a shortage of energy, particularly during peak lactation [12].

Heat stress further complicates this dynamic. High temperatures can reduce the appetite among the cows and change metabolic functions, which leads to less milk production and worse milk quality. It can also affect reproductive efficiency. Thus, effective management strategies, including environmental control, nutritional optimization, and breeding for heat resilience, are vital to sustaining both better health for the cows and their dairy productivity [12]. Figure 1 illustrates a cow's milk yield for approximately two years. The typical lactation cycle pattern can be seen, with a dry period between the two lactations in late 2022.

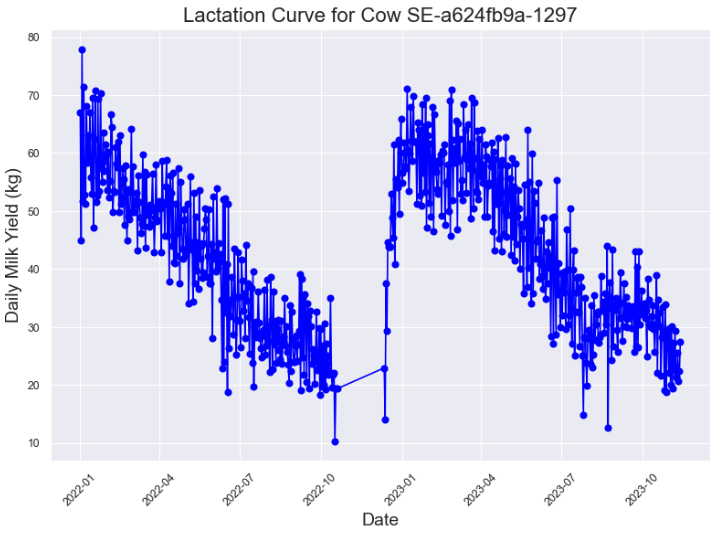


Figure 1: An example of the lactation cycle pattern.

Wood's Lactation Curve

Wood's Lactation Curve is an effective tool in dairy science, offering a mathematical model to characterize and predict milk production over a lactation cycle [13]. Introduced by Wood (1967), this model captures the typical lactation pattern observed in dairy cows through a concise mathematical equation [14]. The equation is formulated as:

$$Y(t) = at^b e^{-ct} \tag{3.1.1}$$

where $Y(t)$ represents the milk yield at time t (days post-calving), and a, b, and c are cow-specific parameters. These parameters are indicative of various aspects of milk production: a correlates with the initial production level post-calving, b governs the incline rate of milk production, and c dictates the decline rate after the peak production [14].

This model's application is an important component in dairy management. By fitting Wood's model to empirical lactation data, one can predict the expected future milk yields [14]. Wood's lactation cycle pattern for the same cow as in Figure 1, is as according to Figure 2.

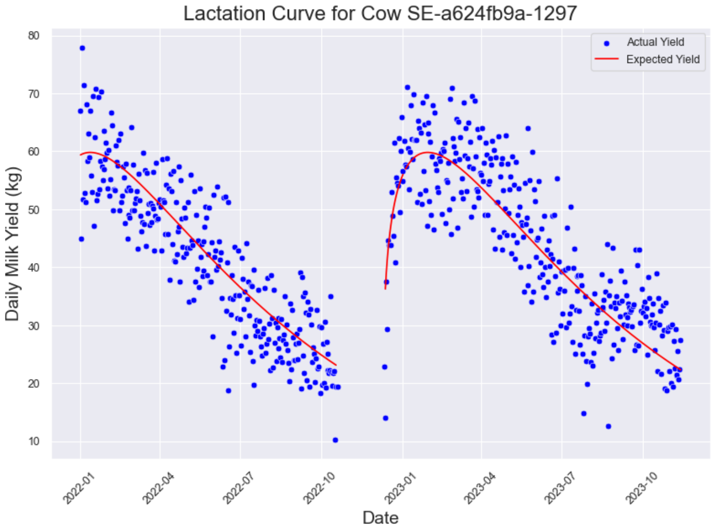


Figure 2: An example of Wood's lactation cycle pattern.

3.2 Analysis of Milk Production Patterns in Dairy Cow Herds

A promising approach to enhance the predictability of milk production leverages the group behaviour of cattle. This methodology is grounded in the hypothesis that dairy cows within a herd exhibit similar production patterns, influenced by both individual and environmental factors [15].

The primary assumption of this approach is that a cow's milk production on a given day (Day X) is likely to be close to its production on the previous day (Day $X - 1$). This consistency forms the baseline for prediction. However, significant deviations in the production of a substantial proportion of the herd might indicate the influence of external factors, such as environmental changes or health issues [15].

The model starts with the simplest form, hypothesizing that each cow should produce approximately the same amount of milk as the day before. This null hypothesis is then tested against actual production data. The statistical significance of deviations from the hypothesis is assessed, providing insights into the factors influencing production. For instance, if a considerable number of cows produce less milk than expected, it might suggest an external factor affecting the herd [15].

Further refinement of the model can include variables such as the stage of lactation of each cow. These refinements aim to enhance the accuracy of the predictions by incorporating more specific individual and temporal factors [15].

3.3 Random Forest

Random Forest, named for its structure resembling a forest made up of many decision trees, is a powerful ensemble learning method used for classification and regression. Each "tree" in this "forest" is a decision-making model, representing a series of choices leading to a final decision. The method builds numerous such trees during its training. For classification tasks, it outputs the mode of the classes (the most common class among individual trees), and for regression, it calculates the mean prediction of the individual trees. This ensemble approach, combining multiple decision trees, significantly enhances the accuracy and resilience of the model [16].

Random Forest creates many smaller groups from the main set of data, each time using different parts of the data. It then builds a decision tree for each group. When making predictions, it combines the results from all these trees. For deciding categories, it picks the most common result from the trees. For numerical predictions, it calculates the average. This approach helps make the model more reliable, reduces the risk of focusing too much on specific data points, and improves its ability to work with new, unseen data [16].

A key mathematical aspect of Random Forest is the reduction of variance in ensemble prediction. Considering a set of N trees, the variance of the ensemble mean can be represented as:

$$\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^{N} T_i\right) = \frac{\sigma^2}{N} + \frac{N-1}{N}\rho\sigma^2 \tag{3.3.1}$$

where T_i is the prediction of the i-th tree, σ^2 is the variance of each tree's prediction, and ρ is the correlation between any two trees in the forest. Given a small correlation, as N increases, the variance of the ensemble mean decreases, leading to more robust predictions if the provided correlation is small [16].

Random Forests also assess feature importance, which is crucial for understanding influential factors in complex datasets. The importance of a feature in Random Forests can be mathematically expressed through the decrease in impurity. In the context of Random Forest and decision tree algorithms, impurity refers to a measure of how well a node in a decision tree splits the data into homogenous subgroups. It's a key concept used to decide how to split the data at each node of the tree. For a given feature f, its importance $I(f)$ is calculated as:

$$I(f) = \sum_{t \in T_f} p(t)\Delta i(s_t, t) \tag{3.3.2}$$

where T_f represents the set of trees where feature f is used for splitting, $p(t)$ is the proportion of samples reaching node t, and $\Delta i(s_t, t)$ is the decrease in impurity due to split s_t at node t [16].

In the context of dairy cow productivity analysis, Random Forest is great at handling complex, non-linear relationships between variables, such as weather conditions and milk production, making it a suitable choice for the datasets being analyzed in this study. Also, Random Forest is less prone to overfitting, which is a significant advantage in this study given the complexity of the data involved [16].

Python Library scikit-learn

Random Forest is a complex mathematical method with a lot of particular details. Fortunately, there are pre-existing built-in functions that simplify its application, which are the ones utilized in this report. Specifically, these built-in functions use the two key mathematical relationships in this report, as illustrated in Equation (3.3.1) and Equation (3.3.2), to effectively implement the Random Forest algorithm without analyzing its computational complexities. '`RandomForestRegressor`' & '`PartialDependenceDisplay`' are the functions from the scikit-learn library used in this study.

The '`RandomForestRegressor`' in scikit-learn is an implementation of the Random Forest algorithm for regression tasks. It builds a multitude of decision trees at training time and outputs the mean prediction of the individual trees. This approach effectively balances the bias-variance tradeoff, making the '`RandomForestRegressor`' robust against overfitting while maintaining a high level of prediction accuracy. This function is particularly efficient in dealing with complex datasets where relationships between variables are non-linear and intricate. For example, in dairy cow productivity analysis, '`RandomForestRegressor`' can easily handle variables such as weather conditions to predict milk production [17].

'`PartialDependenceDisplay`' in scikit-learn is a tool for visualizing the partial dependence of the target variable on a set of 'target' features, marginalizing over the values of all other features (the 'complement' features). In simpler terms, it allows users to see how a target response depends on specific features, holding the other features constant. This functionality is effective for interpreting complex models like Random Forest. It helps in understanding how different features in the data influence the model's predictions. For instance, using '`Partial-DependenceDisplay`' with the Random Forest model in dairy cow productivity analysis can reveal how specific factors like temperature affect milk production, independent of other variables [17].

3.4 Bayesian Linear Regression

Introduction to Linear Regression

Learning the relationship between some variables $\mathbf{X}$ and output $\mathbf{y}$ comes to in linear regression, to learn a model:

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon,$$

where ε is the error term that represents everything about the relationship between $\mathbf{X}$ and the output $\mathbf{y}$ the model failed to capture[18]. The problem is then to find the model $f(\mathbf{X})$ that minimizes ε. The term ε can be viewed as a random variable, independent of $\mathbf{X}$, with a mean value of zero.
We assume that the relationship between $\mathbf{y}$ and $\mathbf{X}$ can be described with a linear function and a constant intercept variable, i.e., an affine function. Hence, if there is p input variables and n data points, size of $\mathbf{X}$ is $n \times (p+1)$ and $\mathbf{y}$'s size is $n \times 1$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{3.4.1}$$

the linear model can be formulated as:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \boldsymbol{\varepsilon} \tag{3.4.2}$$

or:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{3.4.3}$$

The parameter vector $\boldsymbol{\theta}$ will have the size $(p+1) \times 1$ and $\boldsymbol{\varepsilon}$ will be the vector holding all the error terms. Now, let the vector $\hat{\mathbf{y}}$ be defined as $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$, i.e. $\hat{\mathbf{y}}$ is a vector holding the values from the model and not the collected data entries $\mathbf{y}$. From the earlier definition of $\boldsymbol{\varepsilon}$; $\boldsymbol{\varepsilon} = \hat{\mathbf{y}} - \mathbf{y}$, this allows for a *loss function*, $L(\hat{\mathbf{y}}(, \mathbf{y})$, to be formulated, which should indicate how well a model (the choice of $\boldsymbol{\theta}$) fits the observed data $\mathbf{y}$, or how big the vector $\boldsymbol{\varepsilon}$ is. There are plenty of loss functions to choose from, one of the most common is the *squared error loss* and is on the form [18]:

$$L(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2 = \varepsilon_i^2 \qquad i \in \{1, \dots, n\}. \tag{3.4.4}$$

This loss function is 0 when $\hat{y}_i = y_i$ and will grow quadratically when the two differ.
After the choice of a loss function, then a *cost function*, $J(\boldsymbol{\theta})$, is defined as:

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i, y_i) \tag{3.4.5}$$

or the average loss for a given model, over the observed data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ [18]. Finding the most appropriate model then amounts to minimising the cost function:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i(\mathbf{x}_i; \boldsymbol{\theta}), y_i), \text{ where each } \mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}. \tag{3.4.6}$$

With the loss function as the squared error loss, the cost function is called *least squared* cost and can be written in matrix notation as:

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}(\mathbf{x}_i; \boldsymbol{\theta}) - y_i)^2 = \frac{1}{n} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 = \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \tag{3.4.7}$$

To then solve the Optimisation Problem (3.4.6) becomes:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i(\mathbf{x}_i; \boldsymbol{\theta}), y_i) = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \tag{3.4.8}$$

The solution to Equation (3.4.8) has a closed-form solution:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.4.9}$$

when $\mathbf{X}^T\mathbf{X}$ is invertible, which is the case most times [18].

The assumptions that are made are that $\mathbf{y}$ and the error ε are considered random variables, while $\boldsymbol{\theta}$, $\mathbf{X}$ and the variance of ε are considered deterministic [18]. In the following section, these assumptions will slightly change.

Bayes Theorem

Using the methodology in the previous section will yield parameters $\hat{\boldsymbol{\theta}}$ that make the model fit the observed data $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ as good as possible. An alternative methodology is using the *Bayesian approach*, which will instead yield the *distributions* of the parameters conditioned on the observed data $\mathcal{T}$, $P\left(\boldsymbol{\theta}|\mathcal{T}\right)$. This methodology provides more nuance and versatility to the model parameters learned, providing the user with the distributions of the model parameters [18].

Let the probability of an event, denoted A, be $P(A)$, then the probability of both event A and event B happening will be $P(A, B)$, called the *joint distribution* of A and B. Important to emphasise is that event A and event B are regarded as two unrelated events in the previous case, whereas the probability of event A happening given that event B has happened is denoted as $P(A|B)$, called the *conditional probability of A given B* [18].

Let there be a *hypothesis H*, in light of some certain *evidence E*, and the objective is to calculate the probability of said *hypothesis* given the *evidence*. Which is expressed as the *conditional probability of H given E* or $P(H|E)$. To calculate this probability, first, consider how the hypothesis holds before seeing the new evidence, this is known as the *prior* [18]:

$$P(H) \quad - \quad \text{the prior.} \tag{3.4.10}$$

In the case of linear regression, the prior is the user's belief about the parameters' distribution *a-priori*. This often requires some domain expertise or prior experience with the dataset (evidence). Secondly, what is the probability that the evidence would be observed given that the hypothesis holds, known as the *likelihood* [18]:

$$P(E|H) \quad - \quad \text{the likelihood.} \tag{3.4.11}$$

Again with linear regression, this is to say what is the probability of observing the data given the parameters $\boldsymbol{\theta}$? Thirdly, what is simply the probability of observing the evidence without any assumptions or conditions, known as the *marginal likelihood* $P(E)$? Mathematically this can be calculated with the laws of probabilities as[18]:

$$P(E) = \int P(E, H)dH = \int P(E|H)P(H)dH. \tag{3.4.12}$$

Additionally, the marginal likelihood can be viewed as a proportional constant, since it does not scale with the hypothesis [18]. Finally, putting this all together gives *Bayes's theorem*:

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}. \tag{3.4.13}$$

$P(H|E)$ is called the *posterior* [18]. Since the marginal likelihood, can be viewed as a proportional constant when Bayes' Theorem (3.4.13) is viewed as a function of the *hypothesis*, it suffices to just consider:

$$P(H|E) \propto P(H)P(E|H). \tag{3.4.14}$$

This is helpful because the *marginal likelihood* is often expensive to compute [18].

To summarize, in contrast to the classical linear regression introduced in Section 3.4, where the parameters $\boldsymbol{\theta}$ are considered deterministic, in Bayesian Linear regression these are assumed to be random variables, i.e., a *hypothesis* [18].

In contrast to the Least Square method in the previous section, using *the Bayesian approach* lets the user define prior knowledge about the parameters and is beneficial if there is domain expertise suggesting that certain parameters are more likely to be around certain values. Additionally, using *the Bayesian approach* provides naturally the uncertainties of the model quantified, which is not provided in the same manner using the Least Squares method[18].

Assuming that the distributions, and especially the *prior*, are Gaussian distributions will make sense, firstly because many things in nature follow Gaussian distributions, and secondly, setting the *priors* and the *likelihood* of being Gaussian allows for analytical solutions, as will be shown in Section 3.4. The *Bayesian approach* is possible with other assumptions about the underlying distributions but often requires computationally expensive sampling methods[18].

Multivariate Gaussian Distribution

The underlying distribution of the parameters of our model will be assumed to be distributed according to Gaussian distribution, therefore, the methodology will use the multivariate Gaussian distribution frequently.

Let $\mathbf{z}$ be a vector of random variables of size $q \times 1$, all with an underlying Gaussian distribution. The distribution for $\mathbf{z}$ can then be expressed as [18]:

$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\boldsymbol{\mu})\right). \tag{3.4.15}$$

Where $\boldsymbol{\mu}$ is a vector containing the mean values for $\mathbf{z}$ (size $q \times 1$) and $\boldsymbol{\Sigma}$ is the covariance matrix (size $q \times q$) [18]:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \dots & \sigma_q^2 \end{bmatrix}. \tag{3.4.16}$$

The diagonal elements in $\boldsymbol{\Sigma}$ is the variance of the variables in $\mathbf{z}$ and the off-diagonal elements in $\boldsymbol{\Sigma}$ is the covariance between z_i and z_j $(i \neq j)$. The expected value of $\mathbf{z}$ is defined as $\mathbb{E}[\mathbf{z}] = \boldsymbol{\mu}$, the variance $\text{var}(z_i) = \mathbb{E}\left[(z_i - \mathbb{E}[z_i])^2\right] = \sigma_i^2$, and the covariance $\text{cov}(z_i, z_j) = \mathbb{E}\left[(z_i - \mathbb{E}[z_i])\left(z_j - \mathbb{E}[z_j]\right)\right] = \sigma_{ij} = \sigma_{ji}$ $(i \neq j)$ for each i [18].

Bayesian Paradigm in Regression

As in Section 3.4, the model is described as Equation (3.4.3), and ε_i is still assumed to come from a Gaussian distribution with mean zero and variance σ^2 [18].

Likelihood

Model (3.4.3) can now be expressed as a distribution:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}\right), \tag{3.4.17}$$

this can be done because a vector of size $n \times 1$ that has a diagonal covariance matrix, is equivalent to n scalar Gaussian random variables [18]. $\mathbf{I}$ is the identity matrix.

Prior

The prior $P(\boldsymbol{\theta})$, as mentioned in Section 3.4, is postulated by the user, often with domain expert knowledge, in this work assumed to follow a Gaussian distribution:

$$P(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}\right), \tag{3.4.18}$$

where $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$ are the hyperparameters set by the user[18].

Derivation of Posterior Mean & Covariance

Given the *prior* $P(\boldsymbol{\theta})$ and the *likelihood* $P(\mathbf{y}|\boldsymbol{\theta})$, the aim is calculate the *posterior* **mean** and **covariance matrix**. Luckily, there exist two helpful theorems and a subsequent corollary, Theorem A.1, Theorem A.2 and the subsequent Corollary A.2.1, these Theorems and Corollary can be found in the Appendix under section A [18].

To compute the *posterior* distribution then amounts to identifying that in Corollary A.2.1:

$$\mathbf{x}_a = \boldsymbol{\theta}, \quad \mathbf{x}_b = \mathbf{y}, \tag{3.4.19}$$

and thus:

$$P(\mathbf{x}_a) = P(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{0}\right) \text{ (the prior)}, \tag{3.4.20}$$

$$P(\mathbf{x}_b|\mathbf{x}_a) = P(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}; \mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}\right) \text{ (the likelihood)}, \tag{3.4.21}$$

and by using the result from Corollary A.2.1:

$$P(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right) \tag{3.4.22}$$

where:

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}\right), \tag{3.4.23}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}. \tag{3.4.24}$$

This can be implemented and calculated using some basic functions in the Python package Numpy, and be sped up with the JIT-compilator Numba [18, 19, 20].

4 Implementation

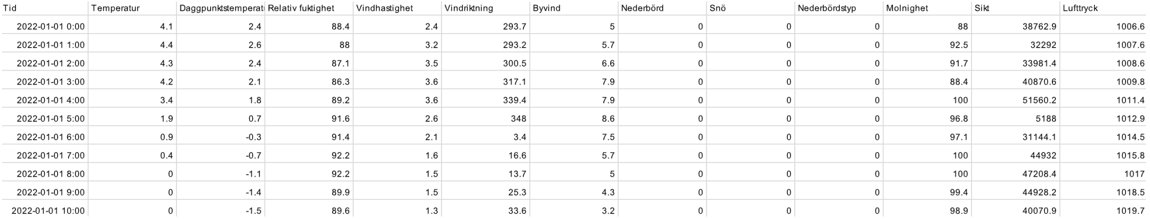
This section covers the practical implementations of the various methods, as well as the data preprocessing applied to the datasets.

4.1 Datasets

This study used two main data sources. The first source was weather data from the Swedish Meteorological and Hydrological Institute (SMHI). The second source was from the SLU infrastructure Gigacow, which gave information about different aspects of dairy cows. By using these two types of data, the study could explore how weather conditions, especially heat, impact dairy cows in terms of both their health and milk production. This was the same sorts of datasets as they studied in last year's report, "The Impact of Weather Factors on The Productivity of Dairy Cows" by Ginlund, R. and Zhou, M., but in this case it's for a different time interval [2].

The Weather Data

The weather data for this project were retrieved from SMHI since the previous reports stated that they have the best historical weather data available for Sweden [2, 1]. The majority of the data was ordered from SMHI by giving them the coordinates for the farms which were of interest in this study, and they delivered it as CSV files, structured in such a way that there was a CSV file named after the farm's pseudonym containing weather data from January 1, 2022, to November 13, 2023. This data came from their MESAN (AROME) analysis model which employs a grid with 2.5km by 2.5km spacing, based on interpolated observations [21]. These CSV files, delivered from SMHI, contained 12 different variables for every timestamp and were formatted according to Figure 3.



Tid	Temperatur	Daggpunktstemperat	Relativ fuktighet	Vindhastighet	Vindriktning	Byvind	Nederbörd	Snö	Nederbördstyp	Molnighet	Sikt	Lufttryck
2022-01-01 0:00	4.1	2.4	88.4	2.4	293.7	5	0	0	0	88	38762.9	1006.6
2022-01-01 1:00	4.4	2.6	88	3.2	293.2	5.7	0	0	0	92.5	32292	1007.6
2022-01-01 2:00	4.3	2.4	87.1	3.5	300.5	6.6	0	0	0	91.7	33981.4	1008.6
2022-01-01 3:00	4.2	2.1	86.3	3.6	317.1	7.9	0	0	0	88.4	40870.6	1009.8
2022-01-01 4:00	3.4	1.8	89.2	3.6	339.4	7.9	0	0	0	100	51560.2	1011.4
2022-01-01 5:00	1.9	0.7	91.6	2.6	348	8.6	0	0	0	96.8	5188	1012.9
2022-01-01 6:00	0.9	-0.3	91.4	2.1	3.4	7.5	0	0	0	97.1	31144.1	1014.5
2022-01-01 7:00	0.4	-0.7	92.2	1.6	16.6	5.7	0	0	0	100	44932	1015.8
2022-01-01 8:00	0	-1.1	92.2	1.5	13.7	5	0	0	0	100	47208.4	1017
2022-01-01 9:00	0	-1.4	89.9	1.5	25.3	4.3	0	0	0	99.4	44928.2	1018.5
2022-01-01 10:00	0	-1.5	89.6	1.3	33.6	3.2	0	0	0	98.9	40070.9	1019.7

Figure 3: The format of the CSV file delivered from SMHI.

While MESAN covers a wide range of weather data, it lacks information on solar radiation. To address this gap, the SMHI STRÅNG model was utilized. This model provides various solar radiation parameters using the same granular grid as MESAN [22]. In this case, only the parameter ”Global Irradiance” was of interest, since it is part of the formula for the Adjusted THI, which leads us into a new topic.

Temperature Humidity Index (THI), recommended by multiple studies, as a better gauge of heat stress in dairy cows, was calculated using the formula for the Mader et al. 2006 adjusted version, named Adjusted THI [23, 24]. The formula for this index is based on the formula presented in Equation (4.1.1).

$$\mathrm{THI_{adj}} = 4.51 + 0.8 \times T + \mathrm{RH} \times (T - 14.4) + 46.4 - (1.992 \times \mathrm{WS}) + (0.0068 \times \mathrm{RAD}) \qquad (4.1.1)$$

In this equation, T represents the dry bulb temperature in degrees Celsius. The variable RH stands for relative humidity, expressed as a percentage. The term WS refers to wind speed, measured in meters per second, and RAD signifies solar radiation, quantified in Watts per square meter. All these variables are given in the SMHI MESAN and STRÅNG files.

A specialized script, mirroring the parameters used in the MESAN script, was developed to consolidate the final, processed, weather dataset. This script performed several functions, including:

1. Loading the delivered MESAN CSV files.
2. Utilizing the STRÅNG API to incorporate solar radiation data.
3. Computing Adjusted THI.
4. Producing new, processed, CSV files that included everything in the delivered MESAN CSV files, Global Irradiance and Adjusted THI.

These newly processed CSV files, containing all weather data of interest in this study, were now ready for further analysis.

The Milking Data

The dataset related to milking comes from the Gigacow project, initiated by SLU. This infrastructure focuses on enhancing communication between researchers and dairy industry professionals. Gigacow's efforts are directed towards developing a platform for gathering data on dairy cows. This platform integrates traditional systems with advanced technologies like new sensors and cameras installed on dairy farms [25].

The milking dataset consisted of five different CSV files. They contained a lot of different variables, e.g., the pseudo-code for the farm names, the ID-code for each cow, the milk yield etc. The variable names of interest in this study and their description can be seen in Table 1.

Table 1: An explanation of the attributes interested for this study in the milking dataset.

Variables	**Description**
FarmName_Pseudo	The pseudo-code for each farm, e.g. a624fb9a.
SE_Number	The cows' ID code, e.g. SE-a624fb9a-1297.
Startdate	The date for the milking record.
LactationNumber	Which lactation cycle the cow is in.
DaysInMilk	Number of days which have gone since the lactation cycle started.
TotalYield	The total milk yield for that record.
SessionNumber	The number of the milking session of the day for that record.
BreedName	The name of the breed.

The first thing to do was to adjust the data to the period of interest in this study, i.e., from January 1, 2022, to November 13, 2023. There were data from 11 different farms in the original dataset. However, upon preprocessing, it became apparent that only two of these farms, `a624fb9a` and `f454e660`, provided adequate data for further analysis for the specified time interval, as can be seen in Figure 4 and Figure 5.

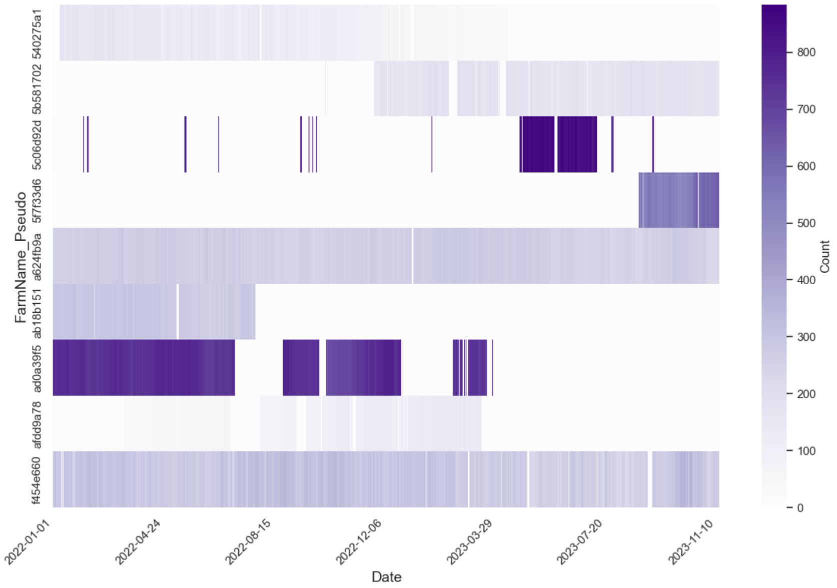


Figure 4: Heatmap of the number of instances of the parameter TotalYield for each farm over time.

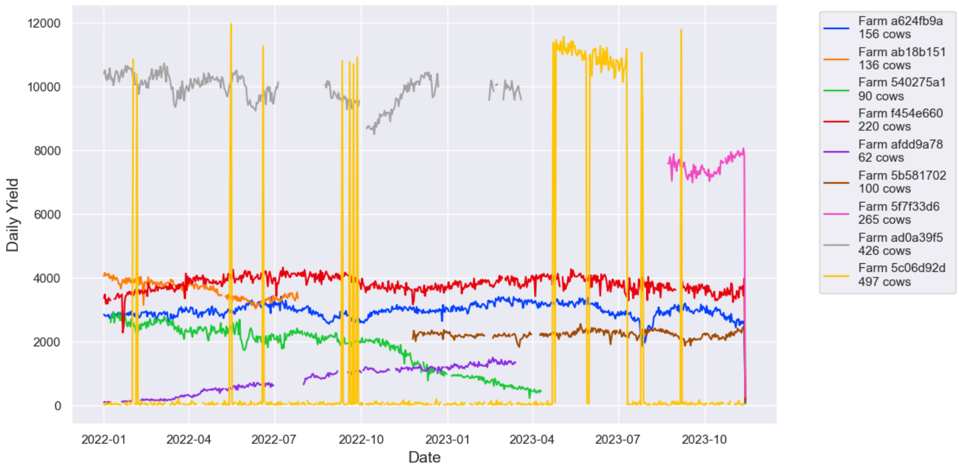


Figure 5: The daily milk yield data for each farm over time.

For those two farms, the pseudo-code of both farms was replaced with ”Farm 1” and ”Farm 2”. `a624fb9a` was renamed to ”Farm 1” and `f454e660` was renamed to ”Farm 2”.

Merging the Two Datasets

In this study, a Python class named ’`MilkDataProcessor`’ was developed to combine the two datasets; the milking dataset and the weather dataset. The functionality of the ’`MilkDataProcessor`’ class include the following:

- *Initialization*: The class is initialized with paths to the milk data and the weather data files, which both have been pre-processed, as described in Section 4.1 and Section 4.1.
- *Loading Milk Data*: It loads the milk yield data, cleans it by removing unnecessary columns like ’`Unnamed: 0`’, ’`LactationInfoSource`’, etc., and converts some columns to appropriate data types. It then merges this data with additional cow information from another file, such as ’`BreedName`’.
- *Adding Weather Data*: Weather data from multiple files is added. This process involves matching the weather data to each farm’s location and dates. The class also identifies heatwave periods, adding the variable ’`HW`’, by analyzing maximum daily temperatures and adding this information to the dataset. According to SMHI, a heatwave is defined as a period of at least five consecutive days with a maximum daily temperature of at least 25 degrees Celsius [26]. The function in question would also check whether there has been a heatwave at any time during the last week. That is to say if there has been the last week or is currently a heatwave, ’`HW`’ is set to 1, and if there has not been a heatwave, then ’`HW`’ is set to 0.

  Further, to differentiate between a heatwave that lasts 5 days and one that lasts, e.g., 8 days, this project introduces a variable named *Cumulative Heatwave* (`cum_HW`). Initially, every heatwave of length 5 is assigned a `cum_HW`$=1$, then for every day a heatwave is ongoing beyond day 5, an increment of +1 is added. E.g. a 6-day heatwave is $[1,1,1,1,1,2]$ and a 7-day heatwave is $[1,1,1,1,1,2,3]$ etc.. Then, in the following week after a heatwave, an exponential decay factor is added, designed to reach 0 after 1 week:

  $$\texttt{cum\_HW}(x_n) = A - 0.01 \cdot \exp\left(x_n \cdot 0.125 \cdot \log\left(100 \cdot A\right)\right), \tag{4.1.2}$$

  where x_n is days after the heatwave ($x_n \in \{1, 2, \ldots, 7\}$) and A is the last value of the heatwave, i.e. for a 7-day heatwave $A = 2$. The pseudo-code for calculating the `cum_HW` is found in the appendix as Algorithm 1.
- *Final Data Preparation*: After merging, the class cleans the combined dataset by removing columns not relevant to the study, such as ’`Snö`’, ’`Molnighet`’, and ’`Sikt`’. It filters the data based on criteria like ’`DaysInMilk`’ and ’`TotalYield`’ to ensure data quality and relevance.
- *Data Extraction Methods*: The class includes methods to extract data for specific farms, cows, or within certain date ranges. This allows for targeted analysis of subsets of the data.

The class is utilized in the main program by creating an instance of '`MilkDataProcessor`' with paths to the milk and weather data. It makes it easy to access all the relevant data shortly and simply, involves loading the milk data, integrating the weather data, and accessing the processed data for analysis.

4.2 Basic Statistics

Exploratory data analysis was conducted to examine the distribution and characteristics of the dataset. Boxplots were utilized to present the average temperature per day for summer 2022 versus summer 2023 and the average milk yield per day for summer 2022 versus summer 2023. These visualizations served to compare the central tendency and variability within each category, highlight potential outliers, and provide a clear depiction of the datasets' five-number summary, which includes the minimum, first quartile, median, third quartile, and maximum values.

4.3 Normalize the Daily Yield Data using Wood's Lactation Curve

This method aims to normalize the daily milk yield data from dairy cows across Farm 1 and Farm 2. This normalization process utilizes Wood's Lactation Curve model to account for the natural variations in milk production over the lactation cycle of each cow.

Wood's Lactation Curve, as described in Section 3.1, provides a mathematical model to characterize and predict milk production over a cow's lactation cycle. The model is expressed as:

$$Y(t) = at^b e^{-ct} \tag{4.3.1}$$

where $Y(t)$ represents the milk yield at time t (days post-calving), and a, b, and c are parameters specific to each cow, depicting different aspects of the milk production curve. The normalization process was as follows:

1. **Parameter Estimation:** For each cow, the parameters a, b, and c could be estimated using Wood's equation based on that cow's historical lactation data. This estimation is crucial for accurately predicting the expected milk yield for each day of the lactation period.

2. **Expected Yield Calculation:** Utilizing the estimated parameters, the expected milk yield $Y(t)$ for each cow for every day of the observed period could be calculated.

3. **Normalization of Actual Yield:** The actual daily milk yield for each cow is then normalized by dividing it by the expected yield for that day. The formula used is:

$$\text{Normalized Daily Yield Ratio} = \frac{\text{Actual Daily Yield}}{\text{Expected Daily Yield}} = \frac{\text{Actual Daily Yield}}{Y(t)} \tag{4.3.2}$$

   This step adjusts the daily yield data, accounting for the natural progression of the lactation cycle.

4. **Data Aggregation and Analysis:** The normalized yields from all cows across both farms are aggregated. This normalized data provides a more standardized basis for comparing milk production patterns and assessing other factors influencing yield, such as environmental conditions.

4.4 Implementation of Random Forest

By normalizing the daily milk yield data or the daily yield change data of each cow on a farm, against the expected output as projected by Wood's lactation curve, a consistent baseline was established for analysis. To unravel the complex, non-linear trends and patterns within this data, Random Forest was employed using Python's free software machine learning library scikit-learn (see Section 3.3). This method constructs multiple decision trees across various variables, thereby enabling a comprehensive non-linear visualization and understanding of the diverse factors influencing milk production, as told in Section 3.3. By building numerous trees for several different parameters, Random Forest was applied to several different approaches.

The Daily Yield

First of all, the daily milk yield for each cow was normalized against expected yields from Wood's lactation curve, setting a uniform standard for analysis. A Random Forest model was then trained on this data, using scikit-learn's '`RandomForestRegressor`' with 1000 estimators (see Section 3.3). This model focused on the mean daily temperature and the mean Temperature-Humidity Index adjusted (THI_{adj}) as input variables. This implementation, adept at capturing non-linear trends, explored the relationships between these environmental factors and normalized milk yields.

The analysis involved creating partial dependence plots, scikit-learn's '`PartialDependenceDisplay`' (see Section 3.3), to visually represent the temperature's and the THI_{adj}'s impacts on milk production. Additionally, the data was segmented based on different heatwave (HW) conditions (0 or 1), allowing for an examination of how these relationships vary under different scenarios.

Moreover, the impact of night temperatures on the daily yield was also examined. Nightly minimum temperatures were categorized as 'Tropical', 'Warm', 'Ideal', and 'Mild', aligning with the categorizations detailed in Section 3.1. By integrating these temperature categories into the Random Forest model, the study could show how different night temperature conditions affect the daily milk yield.

Analysis of Milk Production Patterns in Dairy Cow Herds

In implementing a statistical approach to describe the milk yield, a baseline prediction model is established where the expected milk production of a cow on Day X is set to its production level on Day $X - 1$, as described in Section 3.2.

The next step involved calculating the daily yield change for each cow and normalizing this change against the expected yield from Wood's lactation curve. A Random Forest model was trained with this normalized daily yield change data, using scikit-learn's '`RandomForestRegressor`' with 1000 estimators (see Section 3.3), incorporating the same environmental factors such as for the daily yield.

To visualize the model's findings, partial dependence plots were created, using scikit-learn's '`PartialDependence-Display`' (see Section 3.3). These plots serve as a tool to differentiate between the impacts of mean temperature and THI_{adj} on the normalized daily yield change. Further, the model's capacity to handle complex scenarios was tested by creating conditional plots based on different HW values (0 or 1). The analysis also extended to exploring the influence of night temperatures on the daily yield change. The categorization for the night temperatures was the same as in the section above. By integrating these temperature categories into the Random Forest model, the model could show how different night temperature conditions affect the daily yield change in dairy cows.

4.5 Implementation of Bayesian Linear Regression

Using the framework provided in Section 3.4, there is a large number of possibilities for designing the feature matrix $\mathbf{X}$. In this project, two major different approaches are used and explained in the following Section 4.5 and 4.5.

Linear model

First, a linear model is implemented where the normalized daily yield is the output variable and the model is capable of using a wide range of different parameters as input variables. For the scope of this study, these were used:

- `HW` - heatwave indicator
- `cum_HW` - cumulative heatwave function
- `DaysInMilk` - days in cycle
- `LactationNumber` - cycle number
- `MaxTemperature` - daily max temperature
- `MeanTemperature` - daily mean temperature

- `MinTemperature` - daily min temperature
- `MeanRelHumidity` - daily mean relative humidity
- `MeanTHI_adj` - daily mean THI-adjusted

as well as dummy variables for the different breeds when examining multiple cows. However, the methodology allows for a user-defined set of features.

Suppose a subset of features has been selected ($\mathbf{X}$), and the target variable ($\mathbf{y}$) is the normalized daily yield. The formulation of the model then looks like this:

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \tag{4.5.1}$$

where:

$$f(\mathbf{X}) = \Phi(\mathbf{X})^T \boldsymbol{\theta}, \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{4.5.2}$$

$\Phi(\mathbf{X})$ is a vector of input features:

$$\Phi(\mathbf{X})^T = \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}, \tag{4.5.3}$$

that allows for different features to be used, as well as manipulation of the features, such as adding an offset/intercept term $\phi(x_i)^T = \begin{bmatrix} 1 & x_{i1} & \dots & x_{ip} \end{bmatrix}$ or squaring one term $\phi(x_i)^T = \begin{bmatrix} 1 & x_{i1}^2 & \dots & x_{ip} \end{bmatrix}$.

Then, in Equation (3.4.23) and (3.4.24), simply substitute $\mathbf{X}$ to $\Phi(\mathbf{X})$:

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \Phi(\mathbf{X})^T \mathbf{y} \right), \tag{4.5.4}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \Phi(\mathbf{X})^T \Phi(\mathbf{X}) \tag{4.5.5}$$

to calculate the posterior. In this study, this model is first fitted to one cow and for this purpose, the cow with the most recorded data points is chosen. Subsequently, individual models will be developed for each cow on a farm, followed by the construction of a comprehensive model for an entire farm.

Generalized Additive Model

Secondly, a Generalized additive model (GAM) is applied to the data, also with a Bayesian regression approach. As before, the normalized daily yield is the output variable $\mathbf{y}$ in this model.

GAMs can capture non-linear relationships between the target $\mathbf{y}$ and the selected features $\mathbf{X}$. Flexibility in this model is achieved by dividing the selected features, indicated as $\mathbf{X}$, into separate subsets. Each subset is matched with its basis function. The key task is to determine the best parameter or weight, noted as θ, for these basis functions. This approach allows for the creation of models that are more adaptable, fitting well within the Bayesian framework.

Furthermore, there are many different basis functions to choose from. This study limits itself to examining linear and quadratic basis functions.

The model takes the form:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \theta_k b_k(\mathbf{x}) + \varepsilon = \mathbf{B}(\mathbf{x})^T \boldsymbol{\theta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{4.5.6}$$

In this model, $b(\mathbf{x})$ represents the basis functions, chosen as basis splines (B-splines). These are piecewise polynomial functions, in degrees of 1 (linear) and 2 (quadratic), where each B-spline affects only a small, specific part of the divided feature subspace. The basis dimension, denoted by K, is determined by the number of subsets the selected feature $\mathbf{x}$ is divided into. The matrix $\mathbf{B}(\mathbf{x})$ is then simply:

$$\mathbf{B}(\mathbf{x}) = \begin{bmatrix} b_1(\mathbf{x})^T \\ \vdots \\ b_K(\mathbf{x})^T \end{bmatrix}, \tag{4.5.7}$$

thus, Equation (3.4.23) and (3.4.24) are changed according to:

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \mathbf{B}(\mathbf{x})^T \mathbf{y} \right), \tag{4.5.8}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \mathbf{B}(\mathbf{x})^T \mathbf{B}(\mathbf{x}). \tag{4.5.9}$$

Still, the parameters $\boldsymbol{\theta}$ are initially assumed to have a mean of $\boldsymbol{\mu}_0$ and a covariance matrix $\boldsymbol{\Sigma}_0$.

Moreover, the normalized daily yields are grouped into N temperature bins. Within each bin, temperatures are categorized into N distinct groups. The average yield is then calculated for all data within each bin, and this average becomes the representative value for that temperature bin. Consequently, the data points are effectively reduced to N distinct values.

5 Results

In the next section, we present the results from data preprocessing and model application. The key point from this analysis is that normalizing lactation curves with Wood's lactation model is effective, clearly showcasing yield decreases due to mostly high but also low temperatures.

5.1 Basic Statistics

Before delving into the complex analytical models, it is fundamental to lay the groundwork with an examination of basic statistics. This section provides this foundation by presenting initial insights into the raw data and exploring potential correlations and patterns that could inform more intricate analyses. It provides the essential starting point for further detailed statistical analysis that will come later. Establishing a basic understanding of the data is a crucial first step before moving on to more complex analysis methods.

As told in Section 4.1, the analysis focused on comparing two farms, as the data available for the other farms was insufficient for a comprehensive comparison. This examination began with very basic statistics, contrasting both the temperature conditions during the summers of 2022 and 2023 and the corresponding milk yields observed in these periods.

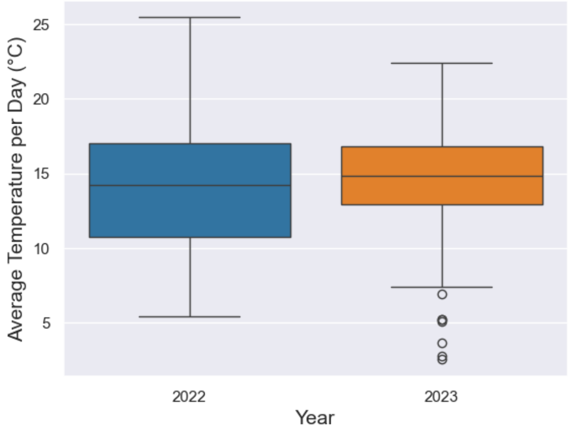


Figure 6: Average daily temperatures for summers of 2022 and 2023 at Farm 1.

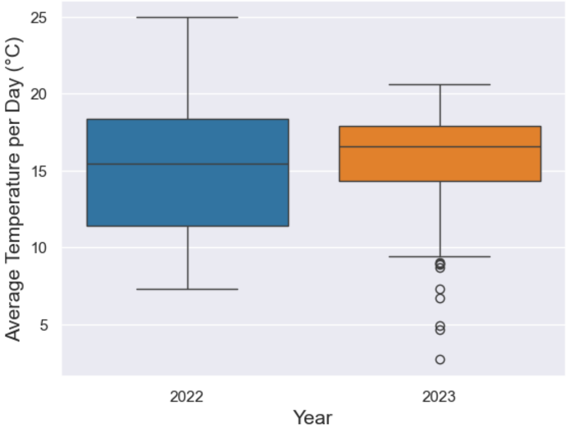


Figure 7: Average daily temperatures for summers of 2022 and 2023 at Farm 2.

The boxplots provide a visual comparison of the average daily temperatures during the summers of 2022 and 2023 for both Farm 1 and Farm 2. For both farms, the average temperatures in the summer of 2023 appear to be higher than in 2022, as indicated by the median line within each box., in Figure 6 and 7. Outliers, marked as individual points, show that there were several unusually cool days in 2023, despite the overall rise in temperatures. These outliers are more prominent in the data for Farm 2. These temperature variations could potentially correlate with the differences in milk yield.

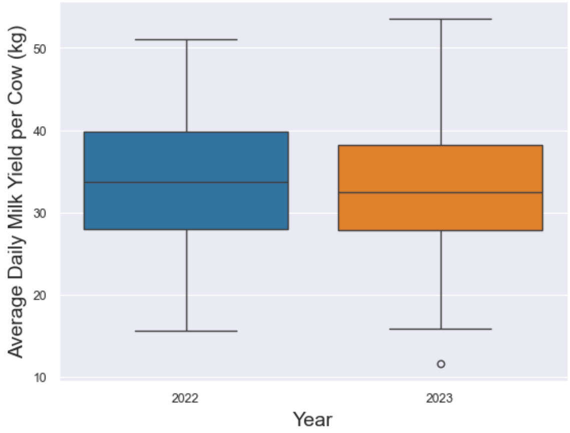


Figure 8: Average daily milk yield per cow for summers of 2022 and 2023 at Farm 1.

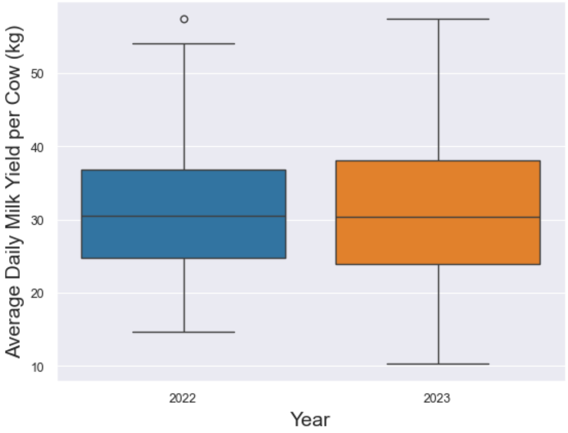


Figure 9: Average daily milk yield per cow for summers of 2022 and 2023 at Farm 2.

The average daily milk yield per cow for both farms shows trends between the summers of 2022 and 2023, as presented in Figure 8 and 9. Farm 1 experienced a slight decrease in median milk yield during the warmer summer of 2023, which aligns with the higher temperatures observed that year. Farm 2's yield remained consistent but with greater variation, potentially reflecting the wider temperature fluctuations seen.

5.2 Results for the Implementation of Random Forest

All the figures in this section, Section 5.2, are partial dependence plots based on the Random Forest models described in Section 4.4 and 4.4.

The Daily Yield

For all the figures in this section, the changes of the normalized daily yield are on the y-axis and it shows how changes in average temperature or average THI$_{adj}$ are predicted to impact the normalized daily yield. The y-axis values represent the magnitude of this impact. For example, a y-value of 1.10 would suggest that at a given temperature, the predicted normalized daily yield is 10% higher than the baseline (where the baseline is the average or expected yield without considering temperature).

The impact on milk production of the average temperature or the average THI$_{adj}$ is presented by the following figures:

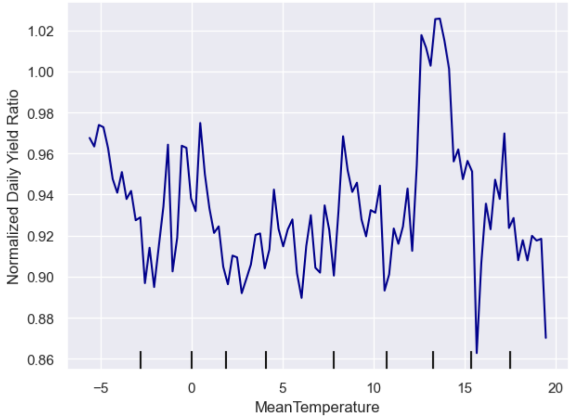


Figure 10: Impact of the mean temperature on daily yield at Farm 1.

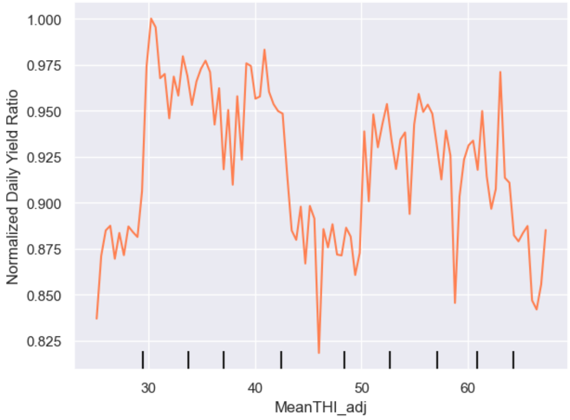


Figure 11: Impact of the mean THI_{adj} on daily yield at Farm 1.

Figure 10 shows yields drop below baseline between -5°C and 15°C, dipping at 0°C, then rising above baseline at 15°C before falling sharply near 20°C, indicating temperature's effect on milk production. Figure 11 depicts a decline in yield with rising THI_{adj} values from 30 to over 60, with a significant drop above 15% under stressful conditions for cows.

The impact on milk production of the average THI_{adj} during non-heatwave (HW=0) days and heatwave (HW=1) days is presented in the following figures:

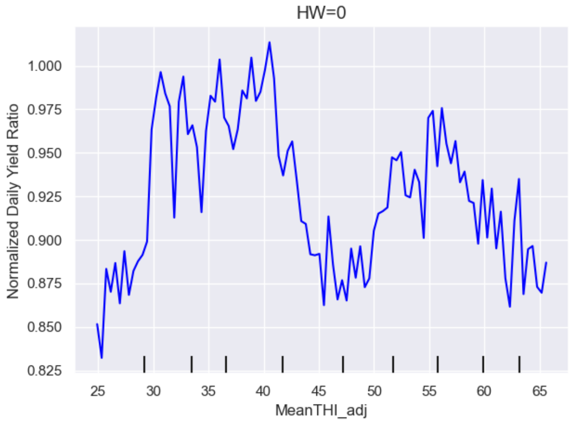


Figure 12: Impact of the mean THI_{adj} on daily yield at Farm 1 for non-HW days.

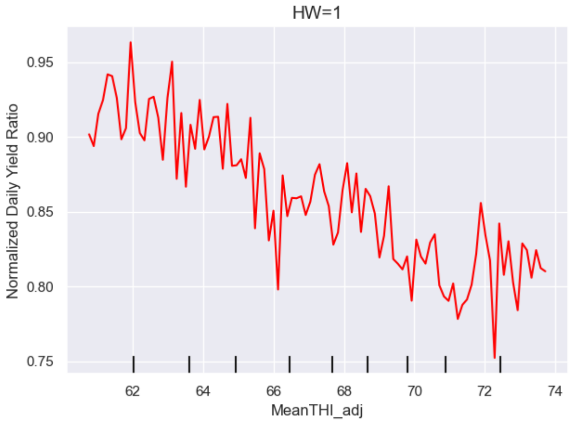


Figure 13: Impact of the mean THI_{adj} on daily yield at Farm 1 for HW days.

Figure 12 shows yield stays near baseline with minor variations and a notable drop of 12.5% at a THIadj of 60 during non-heatwave days. Conversely, Figure 13 shows a distinct drop in yield on heatwave days, with a sharp decrease starting at a THI_{adj} of 62 and down by 20% at 74.

Moreover, the following combined figure showcases the impact of the mean THI_{adj} on the daily yield at Farm 1 across three categories of night temperatures: mild, ideal, and warm:

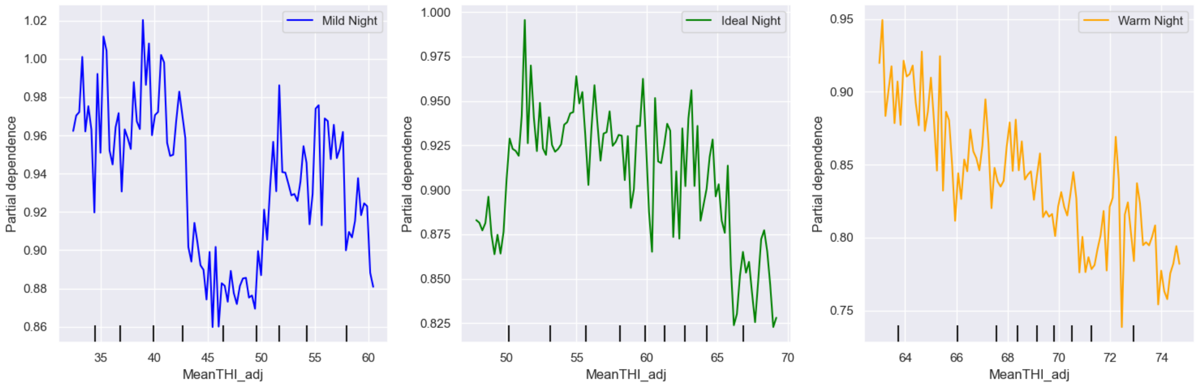


Figure 14: Impact of the mean THI_{adj} on the daily yield at Farm 1 for different night temperature categories.

On mild nights, Figure 14 shows yield fluctuates but stays close to the baseline, dipping slightly as THIadj nears 60. For ideal nights, the yield is more erratic and drops significantly beyond a THIadj of 65. Warm nights see a steady yield decrease, falling up to 20

Analysis of Milk Production Patterns in Dairy Cow Herds

For all the figures in this section, the changes of the normalized yield change are on the y-axis and it shows how changes in average temperature or average THI_{adj} are predicted to impact the day-to-day changes in yield. The primary assumption of this approach is that a cow's milk production on a given day (Day X) is likely to be close to its production on the previous day (Day $X - 1$). This consistency forms the baseline for prediction. Hence, positive y-axis values indicate an increase in yield change compared to the baseline prediction, while negative values indicate a decrease compared to the baseline prediction.

The impact on milk production changes of the average temperature or the average THI_{adj} is presented by the following figures:

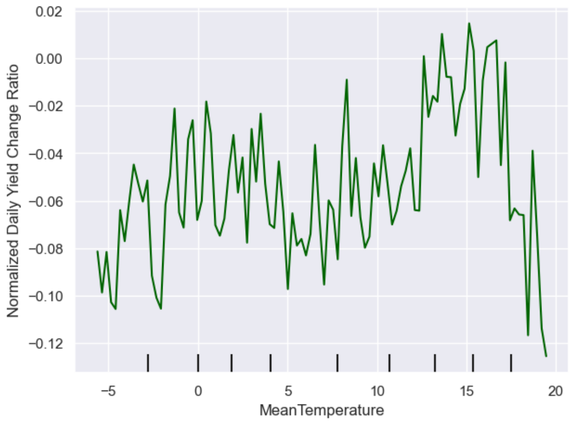


Figure 15: Impact of the mean temperature on the daily yield change at Farm 1.

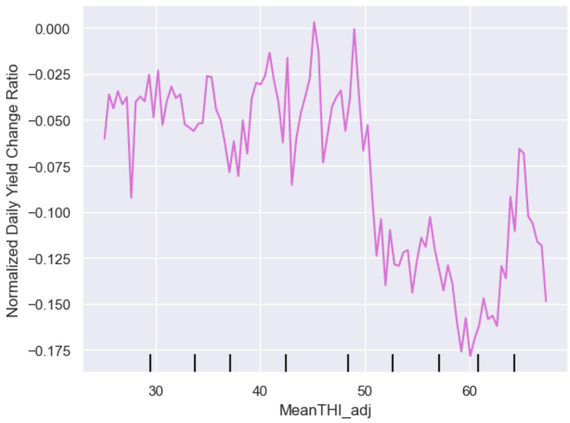


Figure 16: Impact of the mean THI_{adj} on the daily yield change at Farm 1.

Figure 15 indicates a decrease in yield with temperature rise, notably dropping to -10% at 20°C. Figure 16 shows yield declines sharply with higher THI_{adj}, over -15% past 60, highlighting the impact of heat on milk production.

The impact on milk production changes of the average THI_{adj} during non-heatwave (HW=0) days and heatwave (HW=1) days is presented in the following figures:

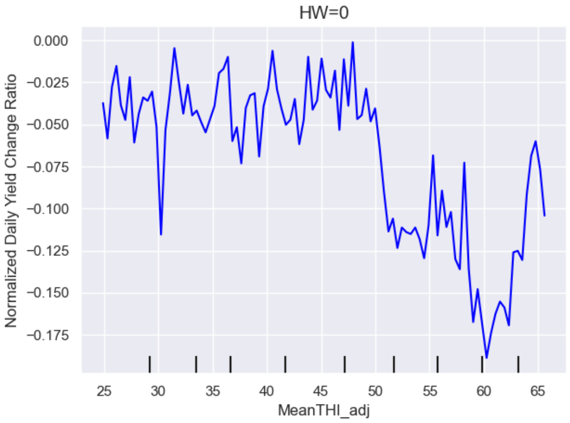


Figure 17: Impact of the mean THI_{adj} on the daily yield change at Farm 1 for non-HW days.

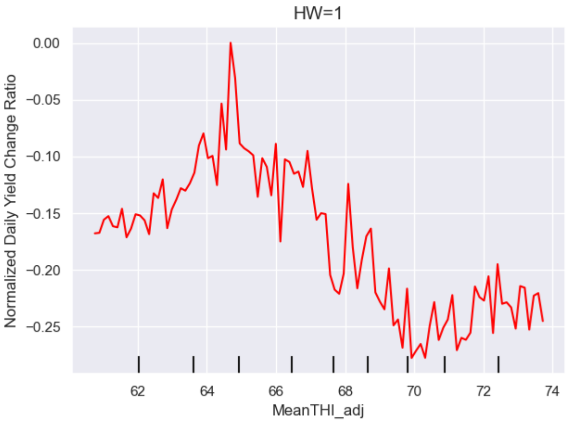


Figure 18: Impact of the mean THI_{adj} on the daily yield change at Farm 1 for HW days.

Figure 17 shows daily yield mostly below zero and falling up to -17.5% with higher THIadj on non-heatwave days. During heatwaves, Figure 18 reveals a steeper yield decline, over -25% as THIadj increases from 62 to 74.

Furthermore, the following combined figure showcases the impact of the mean THI_{adj} on the daily yield changes at Farm 1 across three categories of night temperatures: mild, ideal, and warm:

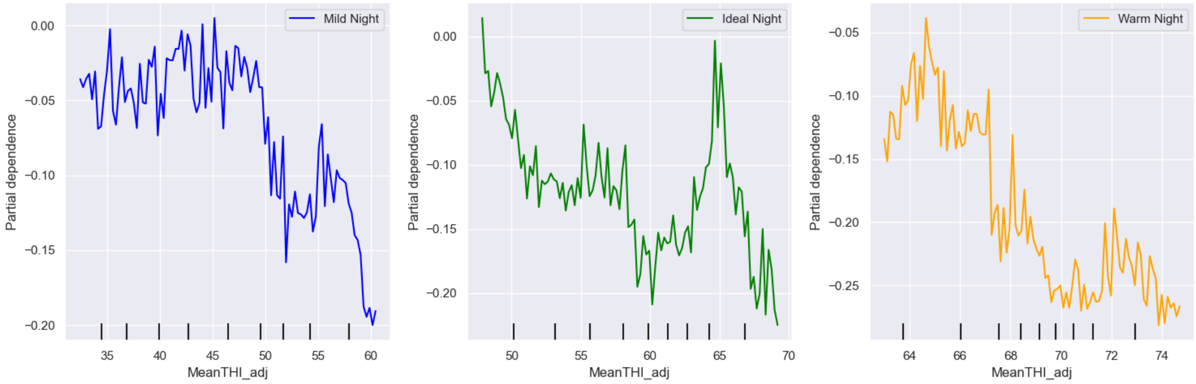


Figure 19: Impact of the mean THI_{adj} on the daily yield change at Farm 1 for different night temperature categories.

Figure 19 shows that mild nights (blue line) see yield dropping to -20% at a THIadj of 60. Ideal nights (green line) have volatile yield changes, with sharp declines past a THIadj of 65. Warm nights (orange line) yield consistently below zero, worsening as THI_{adj} rises, with decreases reaching -25% or more.

5.3 Results for the Implementation of Bayesian Regression

For all Bayesian regression results, a *prior* mean of 0 is used for all parameters and an identity matrix as the covariance matrix, i.e., the variance is set to 1 for all parameters and the covariates to 0. Further, the variance of the *likelihood* is set to the sample variance of the milk yield in the data.

Results Linear Model

This section explores the implementation of the Linear Model, focusing on a range of environmental and physiological features to predict the normalized daily milk yield. The model incorporates variables such as heatwave indicators, temperature extremes, relative humidity, and THI-adjusted values. Initially, the approach involves running individual models for specific cows, capturing the nuances of each animal's response to environmental factors. Subsequent analyses extend to modelling each cow on a farm, and eventually, the entire farm, to understand broader patterns and influences. These varying scales of modelling provide a comprehensive view of how different factors correlate with milk yield in diverse dairy farming scenarios.

One Model for One Cow

Running the linear model suggested in Section 4.5, with the selected features as `HW` (heatwave indicator), `cum_HW` (cumulative heatwave), `MaxTemperature` (daily max temperature), `MinTemperature` (daily min temperature), `MeanTemperature`, `MeanRelHumidity` (daily mean relative humidity) and `MeanTHI_adj` (daily mean THI-adjusted) and choosing the target to be the normalized daily yield for cow `SE-a624fb9a-1297` gives the following results:

Table 2: Posterior distribution for the parameters from model of cow `SE-a624fb9a-1297`

Parameter	$\hat{\mu}$	$\hat{\sigma}^2$
Off-set	0.900	0.120
HW	−0.046	0.100
cum.-HW	0.017	0.078
Daily Max. Temp.	0.032	0.006
Daily Min. Temp.	0.003	0.006
Daily Mean Temp.	−0.029	0.013
Daily Mean Rel. Humidity	0.275	0.050
Mean THI-adj.	−0.005	0.003

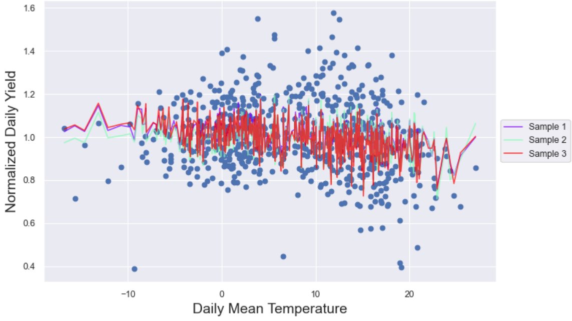


Figure 20: 3 Sampled models from the posterior distribution. Normalized daily yield on the y-axis and daily mean temperature on the x-axis. cow `SE-a624fb9a-1297` *data*

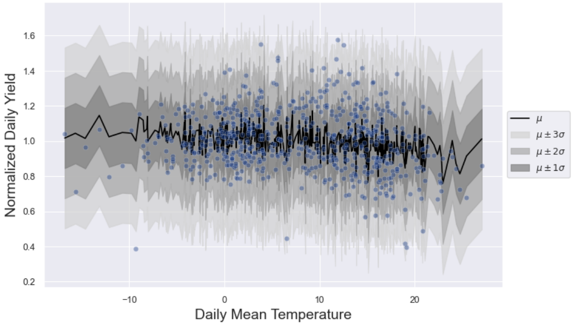


Figure 21: Mean model together with 1 to 3 standard deviations of uncertainty. Normalized daily yield on the y-axis and daily mean temperature on the x-axis. cow `SE-a624fb9a-1297` *data*

One Model for Each Cow on a Farm

Using the same features as in the previous Section 5.3, but adding dummy variables to represent the cow breed, and fitting one model to each cow on Farm 2 yields the following results:

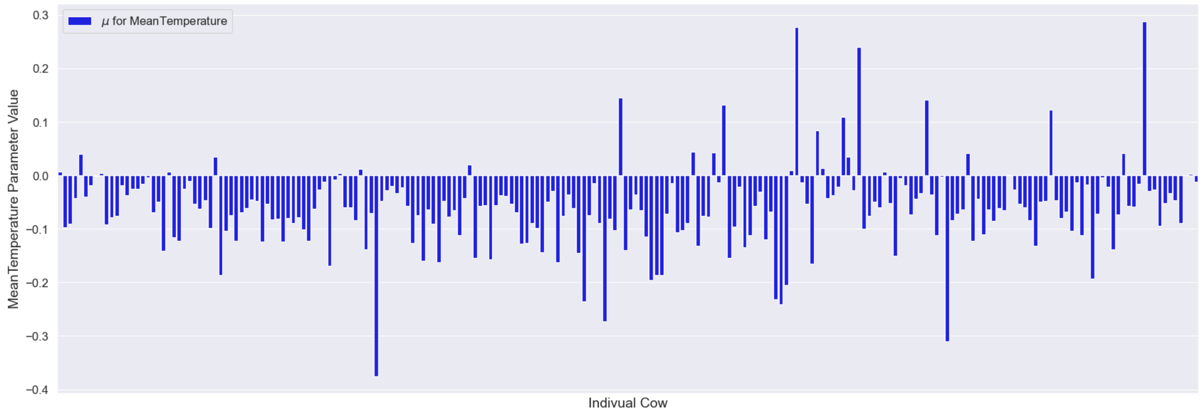


Figure 22: Barplot for each cow's $\hat{\mu}$ for the daily mean temperature parameter (each bar is one cow's mean value). Farm 2.

One Model for Each Entire Farm

Continuing with the same features as in Section 5.3, but now fitting a model to each entire farm yields the following insights about the breeds:

Table 3: Posterior distribution for the breed dummies from the model of Farm 1.

Breed	$\hat{\mu}$	$\hat{\sigma}^2$
SRB	0.9231	0.0157
SLB	0.9196	0.0157

Table 4: Posterior distribution for the breed dummies from the model of Farm 2.

Breed	$\hat{\mu}$	$\hat{\sigma}^2$
SRB	0.7624	0.0102
SLB	0.7603	0.0100
SJB	0.7683	0.0101

The reason for the absence of breed SJB in Table 3 is that there are no cows of breed SJB on Farm 1.

Further, the posterior distributions of the model parameters for the remaining features are visible in Figure 23.

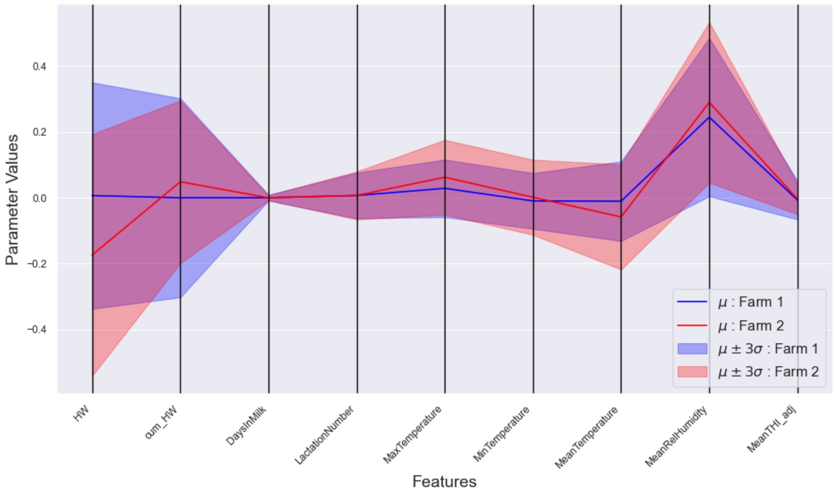


Figure 23: Posterior distributions of the non-breed parameters for two farms. The shaded part is 3 standard deviations from the mean μ.

Results Generalized Additive Model

This section delves into the application of the Generalized Additive Model (GAM), which focuses on daily mean temperature as the sole feature and utilizes the normalized, down-sampled daily yield as the target variable. The analysis is conducted using two different types of basis functions: initially, linear basis functions are employed, followed by an exploration using quadratic basis functions. The results from these distinct approaches offer insights into the relationship between temperature and yield under varying modelling techniques.

Linear Basis Functions

By using linear basis functions, the following results could be visualized:

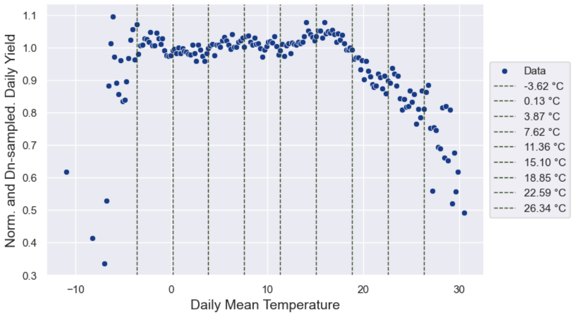


Figure 24: Partitioning of the Farm 2 dataset into 10 *quantiles.*

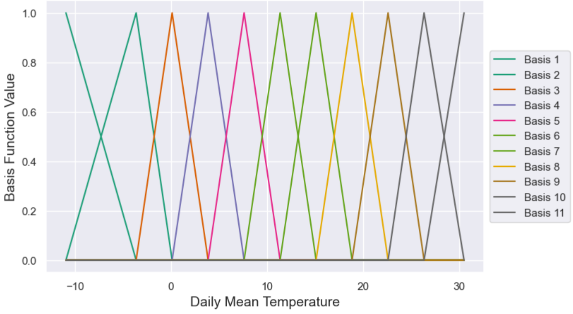


Figure 25: The linear basis functions.

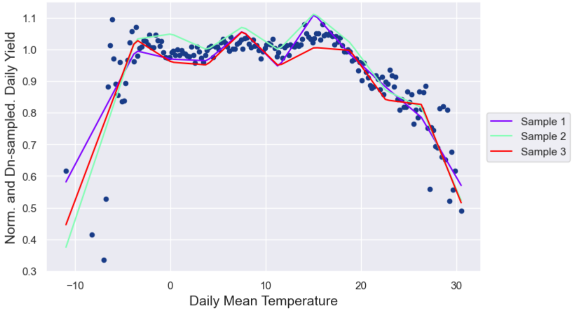


Figure 26: 3 *samples from the posterior distribution. fitted to Farm 2.*

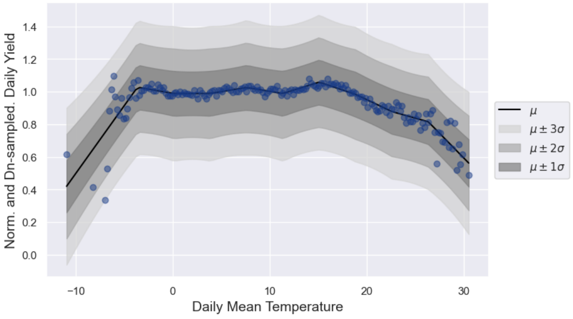


Figure 27: The mean of the distribution with 3 *standard deviations, fitted to Farm 2.*

These results show a plateau of relatively stable yield versus the mean temperature, with cutoffs at −5°C and 15°C respectively for this farm. One of the advantages of using this approach is that it has good interpretability since it becomes clear whether there is an increase or decrease of the yield within the temperature intervals. In addition, by using the *Bayesian approach* the uncertainty of the model parameters is easy to visualise as in Figure 27. To get a higher resolution model, a user could increase the number of quantiles, or use a different partitioning strategy than using quantiles.

Using Least Squares method as mentioned in Section 3.4 and especially Equation (3.4) yields the following result:

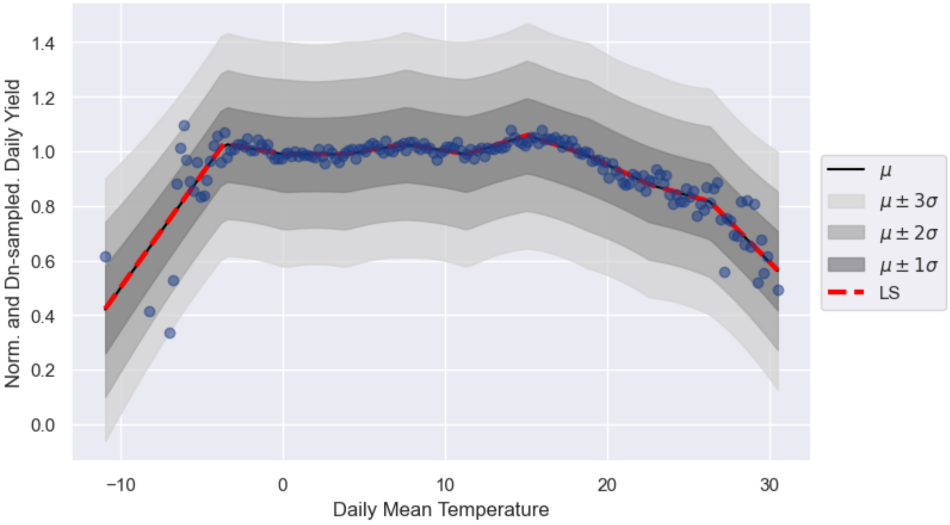


Figure 28: Comparison between Least-Squares method and the Bayesian Approach.

As Figure 28 shows, the mean of the parameters from the *Bayesian approach* and the parameter values from the Least Square method are very similar. However, calculating the sum of the differences between the mean value from *Bayesian approach* and the parameter value from the Least Square method shows that there is a difference:

$$\sum_i (\theta_i^{\mathrm{LS}} - \hat{\mu}_i^{\mathrm{Bayes}}) = 0.012 \tag{5.3.1}$$

Quadratic Basis Functions

By using quadratic basis functions, the following results could be made:

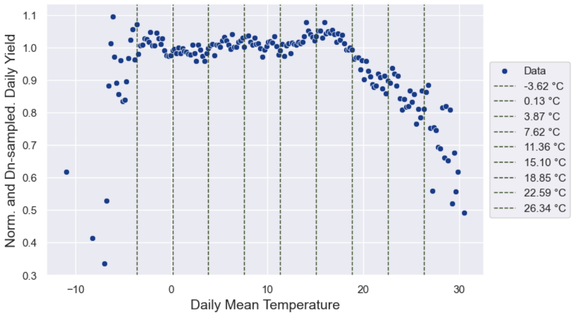


Figure 29: Partitioning of the Farm 2 dataset into 10 *quantiles.*

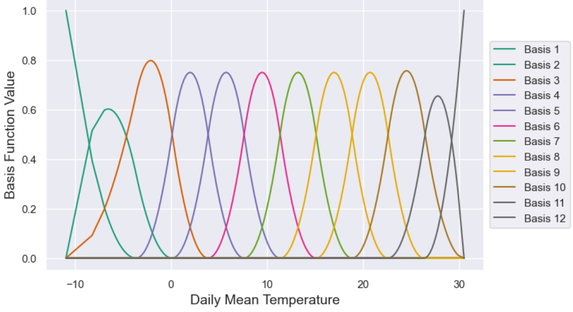


Figure 30: The quadratic basis functions.

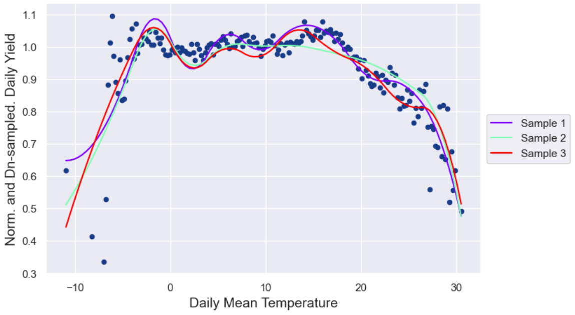


Figure 31: 3 samples from the posterior distribution, fitted to Farm 2.

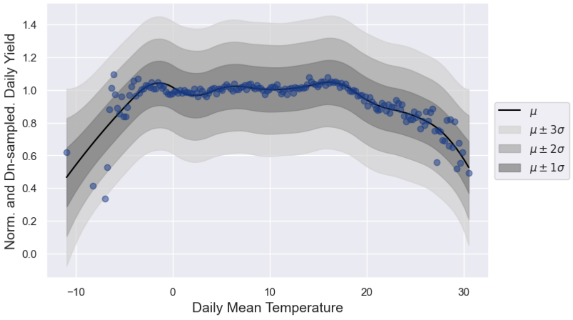


Figure 32: The mean of the distribution with 3 standard deviations, fitted to Farm 2.

Like the results from linear basis functions, using quadratic functions also shows a similar pattern with certain temperatures where changes level off, as seen in Figure 31 and 32. Using quadratic functions is beneficial because they are easy to understand and can pick up more complex patterns in the data. But, there's a downside: these flexible models might fit the data too closely, a problem known as overfitting.

6 Discussion

This section summarizes the findings from the results, followed by a discussion of the results.

6.1 Summary of Results

Basic Statistics

The foundational analysis began with basic statistics, focusing on two farms. Initial insights were gained by comparing temperature conditions and milk yields during the summers of 2022 and 2023. Boxplots revealed higher average temperatures in 2023 than in 2022 for both farms, with Farm 2 experiencing more temperature variability. These results are illustrated in Section 5.1. This initial comparison suggests a potential correlation between temperature variations and differences in milk yield, setting the stage for more detailed analyses.

Random Forest Results

- **Total Daily Yield:** The analysis showed that milk yield varied with temperature changes. At Farm 1, the yield increased up to 2.5 ‰ above the baseline at around 15°C but then decreased sharply as temperatures approached 20°C. High THI_{adj} values led to a significant reduction in yield, particularly on heatwave days. Yield changes were analyzed under different night temperature categories (mild, ideal, and warm). The impact of night temperatures was profound, with warm nights particularly exacerbating the negative effects of high daytime THI_{adj} values on yield. These results are visualized in Section 5.2.
- **Analysis of Milk Production Patterns:** The daily yield change was generally negative with rising temperatures and THI_{adj} values. This trend was more pronounced during heatwave conditions, suggesting a strong negative impact of higher temperatures on daily milk production. As for the results in Total Daily Yield, the study examined yield changes across various categories of night temperatures, specifically classifying them as mild, ideal, and warm. Notably, it was observed that warm night conditions intensified the adverse impacts of elevated daytime THI_{adj} levels on milk production. These results are displayed in Section 5.2.

Bayesian Regression Results

- **Linear Model for One Cow:** Analysis of one cow's data highlighted the influence of various temperature parameters and heatwave indicators on milk yield. The results indicated both positive and negative impacts of different temperature measures on normalized daily yield. These results are depicted in Section 5.3.

- **Linear Model for Each Cow on a Farm:** When analyzing each cow on a farm, significant differences were observed in responses to temperature changes, indicating individual-specific resilience or susceptibility to heat stress. The figure representing these results can be shown in Section 5.3.
- **Linear Model for Entire Farms:** This model focused on the impact of various environmental factors on the overall milk production at the farm level. The analysis revealed differences in dairy production between breeds. These results are visualized in Section 5.3.
- **GAM using Linear Basis Functions for One Cow:** The Generalized Additive Model (GAM) with linear basis functions was applied to analyze the impact of temperature on the daily milk yield of an entire farm. This approach allowed for a very distinct understanding of the relationship between temperature and yield. The model's partitioning into quantiles and its linear basis functions provided a detailed view of how yield varied across different temperature ranges. These results are visualized in Section 5.3.
- **GAM using Quadratic Basis Functions for One Cow:** When applying quadratic basis functions in the GAM for an entire farm, the analysis offered a more nuanced view of the temperature-yield relationship. The quadratic basis functions allowed for the capture of non-linear trends and provided insights into how yield changes at various temperature levels. The model's ability to incorporate more complex relationships between temperature and yield revealed the distinct ways in which environmental conditions impact milk production. The figures representing these results are illustrated in Section 5.3.

6.2 Discussion of Results

The initial analysis using basic statistics has shed light on the possible effects of heat stress on milk yield. However, these basic methods are not enough to capture the full complexity of dairy farm operations. Further analysis, employing more advanced modelling techniques and considering a broader range of variables, is essential to understand the impact of climatic conditions on dairy productivity conclusively.

Temperature and Milk Yield

The observed increase in milk yield up to a certain temperature threshold (around 15°C) at Farm 1, followed by a sharp decline as temperatures approached 20°C, suggests a critical temperature range beyond which cows experience heat stress leading to decreased milk production. This direct relationship between temperature and milk yield suggests a critical temperature threshold for dairy cows, above which heat stress might significantly impact milk production. Furthermore, the influence of night temperatures emerged as a significant factor. The data suggested that warmer night temperatures extended the negative effects of high daytime temperatures on milk yield. This finding shows that not only do daytime temperatures matter, but also the temperatures that cows experience during the night are crucial in determining their overall comfort and productivity.

However, it's important to note that this analysis, based only on temperature, might not capture the complete picture of heat stress on dairy cows, as factors like humidity, wind speed, and solar radiation also play vital roles in determining overall heat stress.

Temperature or $\mathrm{THI}_{\mathrm{adj}}$?

In assessing heat stress among dairy cattle, the decision between using the adjusted Temperature-Humidity Index ($\mathrm{THI}_{\mathrm{adj}}$) and relying solely on ambient temperature (T) becomes a point of discussion.

The $\mathrm{THI}_{\mathrm{adj}}$ formula, incorporating dry bulb temperature (T), relative humidity (RH), wind speed (WS), and solar radiation (RAD) (See Equation (4.1.1)), offers a more complex approach. This complexity could allow a broader analysis of the factors that affect perceived heat, e.g. wind speed can boost cooling, an aspect not captured by only considering temperature alone. However, this complexity also raises questions about its practicality and application, especially in settings where obtaining this data might be a challenge. On the other hand, using just the temperature (T) is simpler, but might not capture the full extent of heat stress, especially in environments where humidity and solar radiation significantly influence perceived heat.

Sometimes, it is rather difficult to explain the complex system we call Mother Earth using mathematics, leading us in onto the next topic.

Wood's Lactation Curve

As mentioned in Section 3.1, the lactation cycle in cows is the period from when a cow gives birth to when it stops producing milk, usually lasting about 305 to 340 days. It begins with a quick rise in milk production right after the cow has a calf. This peaks about six weeks later but then gradually decreases until the cow stops making milk. An efficient model to mathematically describe a cow's lactation cycle is Wood's Lactation Curve model (Equation (3.1.1)). Wood's Lactation Curve model, while simple and effective for analyzing dairy cow milk production patterns, has both strengths and limitations. Its flexibility allows it to adapt to both concave and convex-shaped curves, making it suitable for different breeds and conditions. Additionally, its user-friendly nature makes it easy to implement. On the other hand, its accuracy and effectiveness depend on the availability of comprehensive and high-quality lactation data. In the context of this study, only the farms which had high-quality lactation data were examined. Hence, Wood's Lactation Curve model was very effective for normalizing the daily milk yield data for each cow, making it much more effective to later implement machine learning algorithms such as Random Forest to the dataset.

Random Forest: A Robust Approach for Dairy Data Analysis?

Random Forest is great at capturing complex, non-linear relationships. Its ability to handle multifaceted interactions between variables makes it particularly suitable for this kind of data. One of the key strengths of Random Forest is its resistance to overfitting, especially when dealing with large datasets. This ensures that the models developed are generally reliable and can be applied effectively to predict outcomes. Furthermore, Random Forest can provide insights into the importance of different features (like temperature and THI_{adj}) in predicting milk yield.

All these capacities are important when handling large datasets such as in this study. While Random Forest can handle this complex data well, the resulting models can be quite intricate, making them less interpretable compared to simpler models. Also, the effectiveness of Random Forest models relies on the availability of high-quality data. In cases where data quality is compromised, the model's predictions might be less reliable. Additionally, Random Forest algorithms, especially when dealing with large datasets, can be computationally intensive.

In the case of this study, both of the Random Forest models were made for 1000 estimators, using only the farms which had high-quality data, which took Python approximately 3 minutes each to run. From that, clear visualizations could be made for either the normalized yield data or the normalized yield change versus either the mean temperature or the mean THI_{adj}.

The partial dependence plots showed the impact of these variables relative to a baseline. If the baseline is set by the Random Forest model at an ideal temperature or THI_{adj} range where cows are most productive, any deviation from this range (especially increases) could result in yields that are lower than this optimal baseline. This would explain why the yields are most often seen below the baseline in the figures in Section 5.2.

While the use of Random Forest with a significant number of estimators has provided us with valuable insights into dairy production patterns, it's important to consider the implications of feature selection in our models. This brings us to the discussion on Bayesian regression results and the careful balancing act in the number of features used.

Number of features trade-off

For this project, the number of input features was limited in the various approaches. Mainly because the scope of this project was mainly to asses and provide tools to conduct further research on the matter. Yet, there exist several risks of having too many input features. Firstly, the *Curse of Dimensionality* - having too many input features leads to having a very high dimensional input feature space, and a too-big input feature space can lead to issues with sparsity, meaning that data in the input feature space becomes sparse, requiring large amounts of data to become meaningful. This leads the topic to *overfitting* - requiring large amounts of data due to big input feature space runs a higher chance of noisy data. This will likely lead to any models fitted to the data learning the noise rather than the underlying structures in the data.

Furthermore, increasing the number of features risk of having two or more input features highly correlated. Already with the selected features in this project, the features are probably too correlated, e.g. the daily max, min and mean temperature in the *Bayesian approach*, to give any valuable insights into the parameter values. Additionally, with many input features the model complexity increases making the model harder to interpret which also makes the model harder to visualise, as seen in Figure 20 and 21.

Bayes: Linear model vs GAM model

Bayesian Regression confirms the theory that temperature affects dairy yield, as shown by the negative impact of daily mean temperature in Table 2. Despite this, collinearity among variables suggests a more intricate interaction than initially theorized. The linear model's parameter distributions indicate how features like daily temperatures influence yield and variance, though their high correlation complicates interpretations (Section 5.3). Visuals like Figure 22 help identify individual cows' temperature-yield response, useful for breeding decisions. Differences in breeds' production are highlighted in Tables 3 and 4, while simpler GAM models in Section 5.3 offer clarity by avoiding collinearity and capturing non-linear patterns.

The Bayesian method, using Gaussian priors, provides an analytical solution, but its results are similar to the Least Squares method, as Figure 28 suggests. The advantage of the Bayesian approach lies in providing parameter distributions, offering a nuanced understanding of uncertainties compared to Least Squares' single-point estimates. Exploring other priors like the Gamma or Poisson distributions may reveal complex data patterns not discernible with Least Squares, though this would depart from analytical to computational solutions.

End of Discussion

Concluding the discussion on our study's results, we've delved into the complexities of dairy production analysis, uncovering the intricate relationships between temperature, THI_{adj}, and milk yield. Our approach, combining basic statistical insights with advanced modeling techniques like Random Forest and Bayesian regression, has shown critical aspects of dairy productivity under heated environmental conditions. As a "sanity check" for our models, we identified key temperature thresholds. Our findings consistently indicate that milk production begins to decrease significantly at temperatures around 15°C. Furthermore, for THI_{adj}, we observed an upper threshold of around 60 during heatwaves, and around 65 in their absence. These thresholds align with the expected physiological responses of dairy cows to varying environmental stressors, thereby reinforcing the credibility and applicability of our analytical methods.

7 Conclusions & Further Work

Finally, this section tries to summarize the entire study and provide the major insights, as well as give some suggestions for future work.

The conclusion of this comprehensive study on the impact of Swedish weather conditions on dairy cow health and milk production integrates a range of mathematical and machine learning methods. The study focused on temperature's effects, revealing a critical threshold around 15°C, beyond which milk yield declines sharply, especially during heatwave conditions. Night temperatures were also found to significantly influence milk production.

The analysis began with basic statistical comparisons of temperature and milk yields across various farms and years, hinting at a possible link between temperature fluctuations and milk production changes. It then progressed to applying Wood's Lactation Curve, which normalized the yield data for more complex pattern analysis using machine learning models. The Random Forest method showed promise by revealing intricate relationships between weather conditions and milk yields. Bayesian regression was employed to model data at different levels, from individual cows to entire farms, offering insights despite potential overfitting risks. Lastly, Generalized Additive Models (GAMs), using a Bayesian approach with daily mean temperature as the sole input, provided valuable understandings of the variable relationships and maintained good interpretability due to fewer input features.

In conclusion, this study aimed to offer understanding and resources for examining the heat stress tolerance in dairy cows and their milk production. The integration of Wood's Lactation Curve, Random Forest, and Bayesian regression methods provided a framework for understanding these complex dynamics.

With that said, there is still a lot to do considering heat stress among dairy cows in Sweden. The final aim is to provide insights and tools for further studies that eventually will provide recommendations for farm practices and a decision basis for breeding selection. Hopefully, some conclusions can be made by building upon the results of this study. Some suggestions for further work that can be made from this study are listed below.

Further work

- Investigating the potential of Wood's Lactation cycle with different statistical and machine learning methods, such as neural networks, utilizing weather and cow-specific features to predict deviations from expected lactation cycle values.
- Applying dimension reduction techniques such as PCA, t-SNE and UMAP to the dataset for more efficient data analysis.
- This problem could be extended into a classification problem, labelling individuals based on their heat resilience.
- Reassess the Bayesian framework with different prior and likelihood distributions since the Gaussian prior distributions produce very similar results as LS methods.
- Assess different features and feature combinations. For instance, examine how heat affects different breeds by combining these two features.
- A deeper analysis of the covariance matrix to better understand feature collinearity and interdependencies.

Acknowledgements

First and foremost, we express our gratitude to Stefan Engblom for an outstanding course that was not only enjoyable but also highly educational. However, our deepest appreciation goes to our supervisors: Tomas Klingström, Martin Johnsson, and Lena-Mari Tamminen from the SLU infrastructure Gigacow. Prior to beginning on this study, we had limited knowledge about milk production and related topics, and they provided invaluable assistance, from data support to research insights on heat stress among dairy cows. We would also like to extend our thanks to Wahlsund Säteri, where we spent long days on the countryside, both coding and working on this report. A special mention goes to the cows at Lövsta forskningsstation for being so nice to us during our December visit.

References

[1] R. Hedberg K. Homman and J. Rideg. “Evaluating Locally Measured Weather and Weather Services”. In: *Uppsala University - Project in Computational Science* (2021).

[2] R. Ginlund and M. Zhou. “The impact of weather factors on the productivity of dairy cows”. In: *Uppsala University - Project in Computational Science* (2022).

[3] H. Ahmed, L.M. Tamminen, and U. Emanuelson. “Temperature, productivity, and heat tolerance: Evidence from Swedish dairy production”. In: *Climatic Change* 175.10 (2022), pp. 1–10. DOI: 10.1007/s10584-022-03461-5.

[4] P. Herbut, S. Angrecka, and J. Walczak. “Environmental parameters to assessing of heat stress in dairy cattle—a review”. In: *International Journal of Biometeorology* 62 (2018), pp. 2089–2097. DOI: 10.1007/s00484-018-1629-9.

[5] L. Polsky and M.A.G. von Keyserlingk. “Invited review: Effects of heat stress on dairy cattle welfare”. In: *Journal of Dairy Science* 100 (11 2017), pp. 8645–8657. ISSN: 0022-0302. DOI: 10.3168/jds.2017-12651.

[6] L-M. Tamminen. “Exploring the Impact of Heat and Season on Dairy Cow Health and Fertility – A Mixed Methods Approach”. In: *Proceedings of the Society of Veterinary Epidemiology and Preventive Medicine Conference and Annual Meeting*. Toulouse, France, 2023.

[7] S. Tao and G.E. Dahl. “Invited review: Heat stress effects during late gestation on dry cows and their calves”. In: *Journal of Dairy Science* 96 (7 2013), pp. 4079–4093. ISSN: 0022-0302. DOI: 10.3168/jds.2012-6278.

[8] SMHI. *Tropiska nätter*. https://www.smhi.se/kunskapsbanken/meteorologi/temperatur/tropiska-natter-1.1085. Accessed: 2023-12-19. 2013.

[9] D. Temple et al. “Heat Stress and Efficiency in Dairy Milk Production: A Practical Approach”. In: *FAWEC* (2023). URL: https://www.fawec.org/en/fact-sheets/31-cattle/131-heat-stress-and-efficiency-in-dairy-milk-production-a-practical-approach.

[10] Växa. *Värmestress*. Accessed: 2024-01-06. URL: https://www.vxa.se/fakta/styrning-och-rutiner/mer-om-mjolk/varmestress/.

[11] K. Montgomery. *Sommarhetta Hotar Mjölkförsörjningen*. Extrakt. June 2023. URL: https://www.extrakt.se/mjolkkor-stressas-av-varmen/.

[12] Y. Laurenson E.M. Strucken and G.A. Brockmann. “Go with the flow—biology and genetics of the lactation cycle”. In: *Frontiers in Genetics* 6 (2015). ISSN: 1664-8021. DOI: 10.3389/fgene.2015.00118.

[13] M. Bouallegue and N. M’Hamdi. “Mathematical Modeling of Lactation Curves: A Review of Parametric Models”. In: *Lactation in Farm Animals - Biology, Physiological Basis, Nutritional Requirements, and Modelization*. Ed. by Naceur M’Hamdi. IntechOpen, Dec. 2019. DOI: 10.5772/intechopen.90253.

[14] P.D.P. Wood. “Algebraic Model of the Lactation Curve in Cattle”. In: *Nature* 216 (1967), pp. 164–165. DOI: 10.1038/216164a0.

[15] J. Scheurwater et al. “The effects of cow introductions on milk production and behaviour of the herd measured with sensors”. In: *Journal of Dairy Research* 88.4 (2021), pp. 374–380. DOI: 10.1017/S0022029921000310.

[16] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[17] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[18] A. Lindholm et al. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. URL: https://smlbook.org.

[19] *Numpy documentation*. URL: https://numpy.org/doc/ (visited on 01/04/2024).

[20] *Numba documentation*. URL: https://numba.readthedocs.io/en/stable/ (visited on 01/04/2024).

[21] SMHI. *Analysmodell (MESAN)*. https://www.smhi.se/data/utforskaren-oppna-data/meteorologisk-analysmodell-mesan-arome-api. Accessed: 2023-11-21. 2015.

[22] SMHI. *STRÅNG - en modell för solstrålning*. https://www.smhi.se/forskning/forskningsenheter/meteorologi/strang-en-modell-for-solstralning-1.329. Accessed: 2023-11-21. 2013.

[23] U. Bernabucci et al. “The effects of heat stress in Italian Holstein dairy cattle”. In: *Journal of Dairy Science* 97.1 (2014), pp. 471–486. ISSN: 0022-0302. DOI: https://doi.org/10.3168/jds.2013-6611.

[24] T.L. Mader, M.S. Davis, and T. Brown-Brandl. “Environmental factors influencing heat stress in feedlot cattle”. In: *Journal of animal science* 84.3 (2006), pp. 712–719. DOI: https://doi.org/10.2527/2006.843712x.

[25] SLU. *Gigacow - en SLU-infrastruktur*. https://www.slu.se/institutioner/husdjursgenetik/forskning/gigacow/. Accessed: 2023-11-21. 2021.

[26] SMHI. *Värmebölja*. https://www.smhi.se/kunskapsbanken/meteorologi/temperatur/varmebolja-1.22372. Accessed: 2023-12-11. 2013.

Appendix

A Theory

Here, additional theoretical details are provided to support the closed-form solution to the Bayesian regression problem in Section 3.4 and comes from the literature [18]. They are included for completeness and to offer the reader a more in-depth exploration.

Theorem A.1 (Conditioning)**.** *Partition the Gaussian random vector* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *according to*

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}.$$

The conditional distribution $P(\mathbf{x}_a|\mathbf{x}_b)$ *is then given by*

$$\begin{aligned} P(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \end{aligned}$$

Theorem A.2 (Affine transformation)**.** *Assume that* $\mathbf{x}_a$*, as well as* $\mathbf{x}_b$ *conditioned on* $\mathbf{x}_a$*, are Gaussian distributed according to*

$$\begin{aligned} P(\mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right) \\ P(\mathbf{x}_b|\mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_b; \mathbf{A}\mathbf{x}_a + \mathbf{b}, \boldsymbol{\Sigma}_{b|a}\right). \end{aligned}$$

Then the joint distribution of $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *is*

$$P(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_a \\ \mathbf{A}\boldsymbol{\mu}_b + \mathbf{b} \end{bmatrix}, \mathbf{R}\right)$$

with

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_a\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{b|a} + \mathbf{A}\boldsymbol{\Sigma}_a\mathbf{A}^T \end{bmatrix}.$$

Corollary A.2.1 (Affine transformation - Conditional)**.** *Assume that* $\mathbf{x}_a$*, as well as* $\mathbf{x}_b$ *conditioned on* $\mathbf{x}_a$*, are Gaussian distributed according to*

$$\begin{aligned} P(\mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right) \\ P(\mathbf{x}_b|\mathbf{x}_a) &= \mathcal{N}\left(\mathbf{x}_b; \mathbf{A}\mathbf{x}_a + \mathbf{b}, \boldsymbol{\Sigma}_{b|a}\right). \end{aligned}$$

Then the ***conditional distribution*** *of* $\mathbf{x}_a$ *given* $\mathbf{x}_b$ *is*

$$P(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a; \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right),$$

with

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left(\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a + \mathbf{A}^T\boldsymbol{\Sigma}_{b|a}^{-1}(\mathbf{x}_b - \mathbf{b})\right) \\ \boldsymbol{\Sigma}_{a|b} &= \left(\boldsymbol{\Sigma}_a^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{b|a}^{-1}\mathbf{A}\right)^{-1}. \end{aligned}$$

B Method

The following pseudo-algorithm explains how the cumulative heatwave was calculated and implemented in the weather data.

Algorithm 1: Cumulative Heatwave

```
Data: Max temperatures for each day, $T_1, T_2, \ldots, T_n$
Result: Cumulative heatwave variable, cum_HW
cum_HW ← 0;
Group all consecutive days with daily max temperature ≥ 25°C;
foreach Group of consecutive days with max temperature ≥ 25°C do
    if Number of consecutive days ≥ 5 then
        cum_HW ← cum_HW + 1 ; #Assign to entire group.
        $A \leftarrow 1$;
        if Number of consecutive days > 5 then
            foreach Additional day beyond 5 days do
                $A \leftarrow A + 1$;
                cum_HW ← $A$ ; #Assigned to single day.
            end
        end
        foreach Day $x_t$ in the next 1 week after the heatwave do
            cum_HW ← $A - 0.01 \cdot \exp(x_t \cdot 0.125 \cdot \log(100 \cdot A))$ ; #Assigned to single day.
        end
    end
end
```