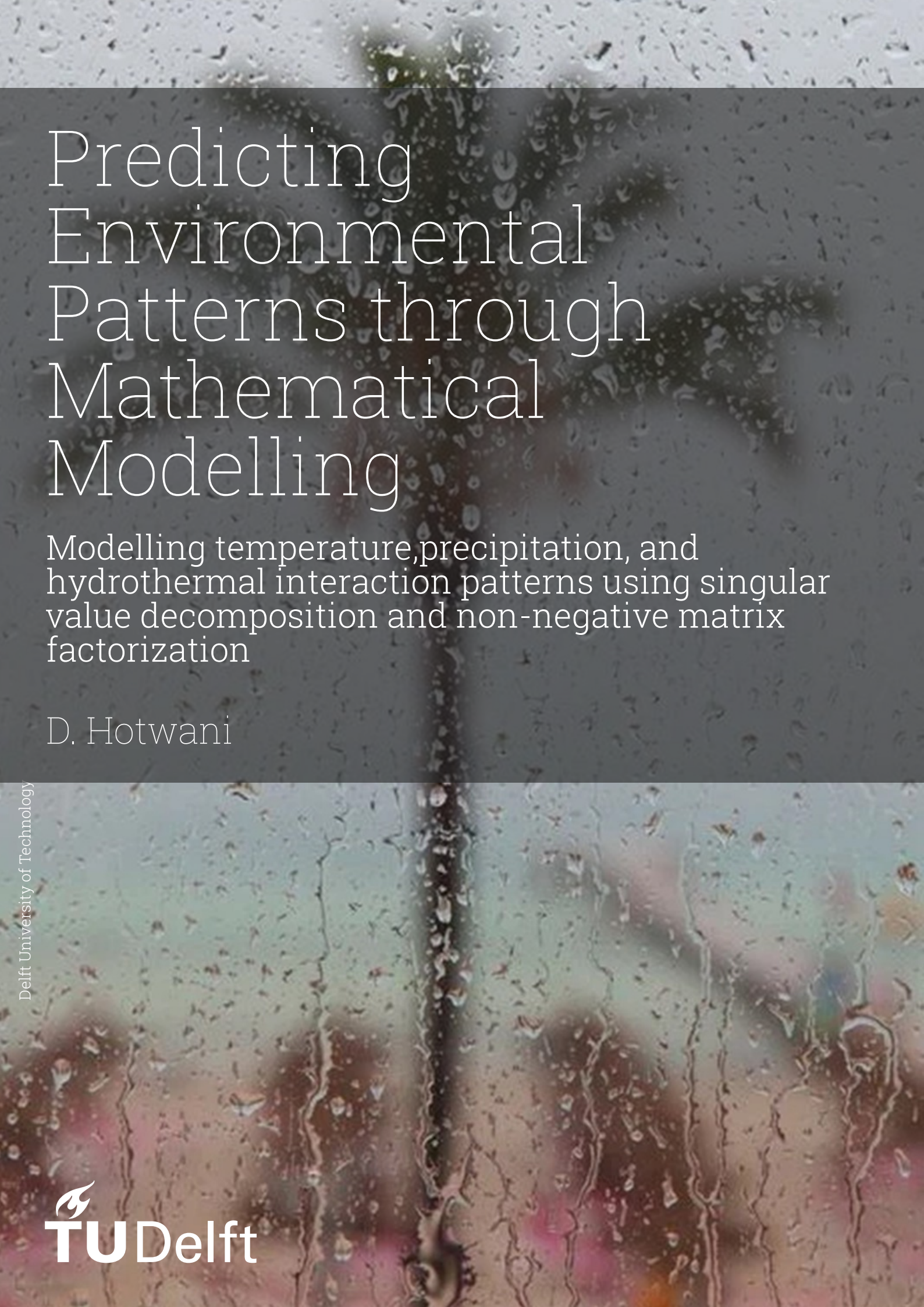

Predicting Environmental Patterns through Mathematical Modelling

Modelling temperature,precipitation, and hydrothermal interaction patterns using singular value decomposition and non-negative matrix factorization

D. Hotwani

Delft University of Technology

TU Delft

Mathematical Insights into Environmental Variable Patterns

Modelling temperature,precipitation, and hydrothermal interaction patterns using singular value decomposition and non-negative matrix factorization

by

Dhania Hotwani

to obtain the degree of Bachelor of Science

at the Delft University of Technology,

to be defended publicly on November 24,2025

Student number: 5247233

Project duration: February 24, 2025 – November 11, 2024

Thesis committee:
Dr. N. Budko, TU Delft, Numerical Analysis, supervisor
Ir. C. Verburg, TU Delft, Numerical Analysis, co-supervisor
Dr. N. Parolya, TU Delft, Statistics

Cover: Modified from Weather Authority Team [1].

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Preface

This thesis is the final project for the Bachelor's degree in Applied Mathematics at Delft University of Technology. In this study, decomposition techniques such as singular value decomposition and non-negative matrix factorization are applied to predict temporal, precipitation, and hydrothermal interaction patterns in the Netherlands.

The choice of topic reflects my deep interest in mathematical modelling and its applications. In 2025, climate in the Netherlands is a subject of considerable debate, and I considered it meaningful to contribute to this discussion from the perspective of a mathematician.

I would like to thank my supervisors, Dr. N. Budko and co-supervisor Ir. C. Verburg, for their guidance and feedback. I also thank the other members of my examination committee, Dr. N. Parolya, for their time and evaluation. Finally, I am deeply grateful to my parents for their unwavering support throughout this process.

Dhania Hotwani
Delft, November 2025

Summary

Earlier studies have applied matrix decomposition methods such as Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF) in climate-agriculture research. These works showed that SVD can reveal interpretable patterns when climate variables display strong, structured signals. NMF's non-negative constraints make it less suitable for representing signed patterns. However, the prediction quality of these approaches has varied widely across regions, variables, and crops, highlighting the need to test their performance in different contexts.

The main research question in this study is: **To what extent do SVD-based and NMF-based representations, in combination with linear prediction models and rank approximations, capture seasonal and interannual variability in temperature, precipitation, and their hydrothermal interactions in the Netherlands?** Both SVD and NMF are applied to the following environment variable data matrices: the temperature matrix, the precipitation matrix, their respective seasonally integrated matrices, and a hydrothermal interaction matrix defined as the integrated product of temperature and precipitation. The seasonal and interannual patterns described by the basis vectors obtained by these decompositions are used to gain insight into the structure of each data matrix. In this thesis, the first components is used to predict seasonal and interannual variations in temperatoral, precipitation, and hydrothermal interaction patterns with a linear prediction model and a rank-approximation. The assessment of these predictions is done by means of the Normalized Mean Squared Error.

The result was that SVD can obtain broad seasonal trends, but these are not always logical in the context of climate behavior. NMF, by contrast, generated positive and interpretable components. The first NMF component consistently represented the dominant seasonal variation, while higher-indexed components resembled residual fluctuations. Overall, NMF provided a clearer and more trustworthy representation of precipitation and hydrothermal interaction patterns, whereas SVD often produced structures that were harder to interpret.

For the prediction of interannual variations, the rank-1 approximation obtained through NMF was generally the most effective. For temperature, NMF rank-1 approximation yielded lower peaks and smaller NMSE values, showing robustness even during extreme years such as 2018 and 2020, and similar improvements were observed for the hydrothermal interaction term. Linear prediction was a simple baseline but proved ineffective. For precipitation, however, prediction remained challenging: both SVD and linear prediction produced high NMSE values, and NMF rank-1 performed similarly to linear prediction, with only temporary improvements in certain years. This highlights that while NMF is the more reliable framework overall, its predictive advantage is not uniform across all variables.

Contents

1

Introduction

The Netherlands experienced warm c combined with lack of rainfall during the years 2018,2020, and 2022 [2–4]. This emphasized the vulnerability of the agricultural systems. It brought attention to adaptive measures, such as adjusting planting and harvesting dates to protect crop yields [5].Stakeholders across research, industry, farming, and policymakers were serious about finding a solution to this problem, as it had occurred beforet. This highlighted the importance of forecasting precipitation,temperature, and the combined hydrothermal interactions.

Previous work has mainly focused on precipitation prediction. For example, radar observations have been combined with deep generative models to forecast rainfall at short lead times in the Netherlands. Their approach was succesful for predicting short-term rainfall at lead times of $5-90$ minutes [6]. These type of approaches are valuable for high-resolution, short-term forecasting.

This study is distinct from earlier works in two ways. First, it delves into three environmental variables: precipitation, temperature and hydrothermal interaction term. The Netherlands relies heavily on water systems such as dikes, drainage, and storage .These protect low lying areas from flooding, and regulate water for the agriculture [7]. Changes in precipitation and temporal patterns, increase the risks of flooding, drainage system failure, altered drought dynamics, and disrupted water supply cycles [8].These combined risks require a framework that treats temperature and precipitation simultaneously. Furthermore, matrix decomposition methods are data-driven techniques, which allow direct identification of dominant seasonal structures.

The motivation of this research can therefore be summarized in three steps. First, we establish an efficient mathematical description of seasonal variations in environmental variables relevant to crop growth: temperature, precipitation, and hydrothermal interaction. Second, we analyze how these variables vary across years, with particular attention to extreme events. Third, we predict future patterns of these environmental variables by using linear prediction. The central research question is: **To what extent do SVD-based and NMF-based representations, in combination with linear prediction models and rank approximations, capture seasonal and interannual variability in temperature, precipitation, and their hydrothermal interactions in the Netherlands?** To address this research question, the following sub-questions were formulated:

- Which modes/components are obtained by applying decomposition techniques to the different data matrices?
- How can the obtained modes/components be used for predicting the Netherlands' temperature, precipitation, and hydrothermal interaction patterns?
- What is the quality of predictions when these modes/components are used to predict temperature, precipitation ,and hydrothermal interaction patterns?

First, the preprocessing of the temperature, precipitation, and hydrothermal interaction data is described in chapter 2, followed by a discussion of the analytical methods in chapter 3. In chapter 4,

the temperature data are analyzed using singular value decomposition and non-negative matrix factorization, which enables a direct comparison between orthogonal and non-orthogonal decompositions in temporal context. Chapter 5 examines precipitation dynamics, evaluating how different representations capture long-term precipitation trends in the Netherlands. In chapter 6, hydrothermal interactions are studied, as they combine two ever-changing environmental variables. Finally, chapter 7 synthesizes the findings to answer the research question.

2

Data description

This chapter provides an overview of the datasets used in the study and outlines the preprocessing procedures applied to them. In addition, it clarifies the methodological choice to use relative humidity data as an approximation for precipitation.

2.1. Data sources, temporal coverage, and variable choice

This section describes the data sources used in the study and explains the reason for using two distinct datasets, including the choice to rely on relative humidity data as a proxy for precipitation rather than the precipitation dataset itself. In addition, the temporal coverage and the selected analysis window are introduced.

As mentioned earlier, different datasets are used for temperature and precipitation variables. For the temperature variable, data from NASA OSDR is utilized. We use the gridded, quality controlled 2m air temperature from NASA's Open Source Data Repository (OSDR), which is accessed via its RESTful API [9]. The benefits of using this source are: offers entire coverage of the Netherlands which reduces representatitivness and limites reproducibility outside the Netherlands, descriptive information about the dataset such as variable names, units and formats. These factors make it easier to work with the data, and to calculate accumulated temperature metrics.

For moisture conditions, the relative humidity (RH) dataset from KNMI De Bilt is used. This dataset offers a long and consistent record at $2m$ height. While precipitation reflects event-based rainfall, RH captures the continuous state of atmospheric moisture. This makes it suitable for trend analysis. Differences in instrumentation and site exposure mean that precipitation sources differ from those used for temperature. In this study, NASA temperature data are combined with KNMI moisture records as a deliberate choice. Both datasets report near-surface conditions at the standard meteorological screen height $(2m)$, ensuring comparability between sources.

The datasets differ in size. This influences the predictive analysis in this study which is based on the $80\% - 20\%$ rule. The temperature dataset starts from April 1, 1981 to October 31, 2024. While, the precipitation dataset starts from January 1, 1930 to October 31, 2024. For the hydrothermal interaction dataset size, we restrict the core analysis to 1981–2024, the full overlap between both datasets.

2.2. Temperature Data Matrix

Daily 2 m air temperature for April–October 1981–2024 was obtained programmatically via the NASA OSDR API [9]. Data are parsed into a Pandas dataframe, with one row per day and columns corresponding to day-of-year within each season as shown in equation (2.1). In this matrix, the first and last five rows of the temperature matrix are shown. The temperature matrix contains 44 rows, each referring to the respective year. Furthermore, there are 214 columns in this matrix, each representing a day of the season. There are no negative entries in this matrix, which is explained by the oceanic climate in the Netherlands. This implies that the average daily mean temperatures from April to October stay

above freezing (greater than 0 degrees Celsius)[10].

$$\underbrace{\begin{bmatrix} 8.70 & 8.24 & 8.13 & 6.71 & 6.96 & \dots & 9.21 & 11.58 \\ 6.56 & 7.94 & 7.74 & 8.83 & 10.15 & \dots & 7.97 & 8.46 \\ 6.87 & 5.81 & 3.95 & 3.02 & 4.98 & \dots & 7.61 & 10.46 \\ 2.84 & 3.26 & 3.00 & 3.70 & 3.60 & \dots & 12.04 & 10.67 \\ 10.36 & 8.44 & 10.99 & 11.92 & 10.29 & \dots & 5.17 & 3.38 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 4.37 & 6.95 & 6.70 & 7.54 & 10.07 & \dots & 13.85 & 13.02 \\ 8.16 & 5.90 & 6.44 & 6.01 & 4.66 & \dots & 12.16 & 11.64 \\ 3.14 & 3.67 & 4.64 & 7.39 & 9.76 & \dots & 15.45 & 13.96 \\ 8.75 & 5.95 & 5.04 & 5.15 & 6.36 & \dots & 11.37 & 11.94 \\ 9.17 & 9.55 & 10.73 & 11.26 & 12.52 & \dots & 13.14 & 12.75 \end{bmatrix}}_{214 \text{ columns (days)}} \Bigg\} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} 44 \text{ rows (years)} \tag{2.1}$$

2.3. Precipitation Data Matrix

Precipitation does not capture the day–to–day atmospheric moisture conditions between rainfall events. RH, by contrast: determines evaporative demand through its role in vapor pressure deficit and thus in evapotranspiration rates. It also influences plant physiology and disease pressure. This affectis stomatal behaviour, latent cooling, and conditions conducive to fungal outbreaks [11]. Furthermore, RH data represents microclimatic stress drivers better than rainfall totals, aligning with the crop–stress modelling framework of this study.We therefore focus on RH for the main analysis. This is used as an approximation for precipitation.

Daily RH from KNMI De Bilt [12] is downloaded for January 1930–October 2024, filtered to April–October for each year. The dataframe has the same structure as the temperature dataset. The 30 missing values were replaced with 0% RH. This is not physically possible for De Bilt and can lead to strong outliers. Doing this has many potential impacts such as: extremely low outliers which drag down seasonal averages, and unrealistic extremes which can distort relationships when modelling. So, RH gaps are better addressed with short–gap interpolation for longer runs. So, very large gaps which are missing are left out for analyses.

In equation (2.2), the first and last few rows of the precipitation matrix are displayed. We observe that there are no negative entries, as negative precipitation does not exist. Furthermore, this matrix consists of 95 rows, referring to the years taken into account. It also consists of 214 columns, referring to the days in the season taken into account.

$$\underbrace{\begin{bmatrix} 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & \dots & 11.0 & 6.0 \\ 0.0 & 0.0 & 0.0 & 17.0 & 48.0 & \dots & 110.0 & 8.0 \\ 14.0 & 54.0 & 18.0 & 2.0 & 1.0 & \dots & 231.0 & 82.0 \\ 27.0 & 3.0 & 14.0 & 0.0 & 0.0 & \dots & 43.0 & 62.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \dots & 51.0 & 25.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \dots & 26.0 & 27.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 16.0 & \dots & 159.0 & 155.0 \\ 49.0 & 0.0 & 0.0 & 207.0 & 32.0 & \dots & 0.0 & 1.0 \\ 132.0 & 0.0 & 0.0 & 0.0 & 0.0 & \dots & 168.0 & 185.0 \\ 0.0 & 14.0 & 70.0 & 120.0 & 30.0 & \dots & 7.0 & 0.0 \end{bmatrix}}_{214 \text{ columns (days)}} \Bigg\} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} 95 \text{ rows (years)} \tag{2.2}$$

2.4. Hydrothermal Interaction Data Matrix

So far, we have reviewed the temperature ,and the precipitation data matrices. In this section, we examine the hydrothermal interaction data matrix. First, we give a motivation for looking at this variable.

Then the layout of the matrix will be explained.

The hydrothermal interaction is the effect when temperature and precipitation are combined. This is a key factor in determining the crop yield. The following scenario highlights the importance of this term: Imagine that the temperature rises. This leads to increased evapotranspiration and water demand. Furthermore, assume that precipitation is low, which means that there is no rainfall. Then, plants suffer from a lack of water, and their photosynthesis is reduced; they become less fertile and grow too quickly. These circumstances lead to great yield reductions.

The dataset for the hydrothermal interaction variable was obtained by using the datasets of the temperature and the precipitation variable. Consequently, the data matrix consists of 44 rows, and 214 columns.

3

Methods

3.1. Dynamic crop growth models

A general Dynamic Crop Growth Model (D-CGM) can be expressed as the following ordinary differential equation:

$$\frac{dY}{dt} = R(v_1, v_2, \dots), \quad Y(t_0) = Y_0 \tag{3.1}$$

where $Y(t)$ is the plant phenotype, e.g., the weight of storage organs, and R is the growth rate function depending on the environmental variables $v_i(t)$, $i = 1, 2, \dots$, such as temperature, soil moisture, evapotranspiration etc. Recently [13], a class of D-CGM's was suggested where the growth-rate function is expressend as a polynomial in environmental variables:

$$R(v_1, v_2, \dots) = \alpha_1 + \alpha_2 v_1 + \alpha_3 v_3 + \alpha_4 v_1^2 + \alpha_5 v_1 v_2 + \dots \tag{3.2}$$

This leads to the following explicit solution of (3.1):

$$Y(t) = Y_0 + \sum_{i=1}^{n} \alpha_i w_i(t), \tag{3.3}$$

where the functions $w_i(t)$, $i = 1, \dots, n$ are the cumulative integrals:

$$\begin{aligned} w_1(t) &= \int_{t_0}^{t} dt' = t - t_0 \\ w_2(t) &= \int_{t_0}^{t} v_1(t')\, dt', \\ w_3(t) &= \int_{t_0}^{t} v_2(t')\, dt', \\ w_4(t) &= \int_{t_0}^{t} v_1^2(t')\, dt', \\ w_5(t) &= \int_{t_0}^{t} v_1(t') v_2(t')\, dt', \\ &\dots \end{aligned} \tag{3.4}$$

Predicting the growth of a crop using this model requires the knowledge of the cumulative integrals (3.4) for the upcoming season. One way to achieve this is to predict the environmental variables $v_i(t)$ and then compute the cumulative integrals of the predictions. On the other hand, one could focus on predicting the cumulative integrals $w_i(t)$ themselves, without first predicting the environmental variables.

3.2. Estimating cumulative data integrals

In this section, we estimate cumulative data integrals of the environmental variables: temperature, precipitation, and hydrothermal interaction. We do this as these integrals appear in growth models such as WOFOST [14]. From plant physiology, we know that the environmental variables, such as temperature, influence the instantaneous rate at which plants grow. So, the size of the plant at a certain time t often depends on the cumulative integrals of the environmental variables [15].

Climate databases typically contain the daily averages of environmental variables. By definition, the daily average $\overline{q}_i$ of the variable $q(t)$ is given by:

$$\overline{q}_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} q(t)\, dt. \tag{3.5}$$

The cumulative integral $S_q(t)$ of $q(t)$ from t_0 to t is defined as:

$$S_q(t) = \int_{t_0}^{t} q(t')\, dt'. \tag{3.6}$$

While computing this integral for an arbitrary time point t can only be done approximately, it is easy to show that the values $S_q(t_i)$, corresponding to the individual days within the season can be obtained from the daily averages exactly.

It is clear that the integral of $q(t)$ over the day that starts at t_{i-1} and ends at t_i can be obtained from the daily average as follows:

$$\int_{t_{i-1}}^{t_i} q(t)\, dt = (t_i - t_{i-1})\overline{q}_i. \tag{3.7}$$

Therefore, the cumulative integral $S_q(t_i)$ on the day that ends at t_i equals

$$S_q(t_i) = \sum_{j=1}^{i} \int_{t_{j-1}}^{t_j} q(t)\, dt = \sum_{j=1}^{i} (t_j - t_{j-1})\overline{q}_j, \tag{3.8}$$

and can be directly obtained from the daily averages.

As previously mentioned, the season that we are interested in for this research starts on the 1st of April at 0:00h, denoted by t_0. It ends on the 31st of October at 23:59h, denoted by t_{214}. So we are considering the interval $[t_0, t_{214}]$, with time step $\Delta t = 1$ day.

Let the daily average data $\overline{q}_i$, $i = 1, \ldots, 214$ for the variabloe $q(t)$ be stored in the vector $\mathbf{w}_q \in \mathbb{R}^n$, $n = 214$. Further, we define the lower-triangular matrix $C \in \mathbb{R}^{n \times n}$ with the elements:

$$[C]_{ij} = \begin{cases} 1, & i \geq j, \\ 0, & i < j. \end{cases} \tag{3.9}$$

Then, the vector $\mathbf{s}_q \in \mathbb{R}^n$, containing the daily values of the cumulative integral $S_q(t_i)$, $i = 1, \ldots, n$, can be computed as

$$\mathbf{s}_q = C\, \mathbf{w}_q. \tag{3.10}$$

If the Environmental Variable (EV) of the type q is collected over m years, the corresponding data rows $\mathbf{w}_q^T$ will be stored in the matrix $W_q \in \mathbb{R}^{m \times n}$, and the matrix $S_q \in \mathbb{R}^{m \times n}$ with the rows $\mathbf{s}_q^T$ of the cumulative integrals will be obtained as $S_q = W_q C^T$.

3.3. Singular Value Decomposition

In this study, we aim to find a basis that captures both seasonal and interannual variations of the environmental variables: temperature, precipitation, and hydrothermal interaction. One of the methods used for this is the singular value decomposition (SVD). Our choice for using this technique is justified

by the data-driven aspect of this method, which extracts dominant spatio-temporal patterns from the data itself. It also automatically ranks them by size of the singular values. This enables interpretable reconstructions.

The following theorem states the existence of the SVD for any data matrix:

Theorem: Let A be an $m \times n$ real matrix. Then $A = U\Sigma V^T$. Here U is an $m \times m$ orthogonal matrix. V is an $n \times n$ orthogonal matrix, and Σ is an $m \times n$ diagonal matrix with r non-zero entries $\sigma_k > 0$, $k = 1, \ldots, r$, where $r = \mathrm{rank}(A)$. These entries are called singular values.

In the singular value decomposition

$$A = U\Sigma V^T = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T, \tag{3.11}$$

to each index $k = 1, 2, \ldots, r$ there corresponds a pair of singular vectors: the k-th left singular vector $\mathbf{u}_k$ and the k-th right singular vector $\mathbf{v}_k$ [16].

The matrices $U,$ and V^T contain the following information:

- The right singular vectors, i.e., the columns $\mathbf{v}_k \in \mathbb{R}^n$, $k = 1, \ldots, r$ of the matrix V, are the seasonal patterns of the environmental variable detected by the SVD algorithm. From these right singular vectors, we can sometimes infer whether it is an early/late start or end of the season by simply interpreting the peaks and dips during the season.
- The left singular vectors, i.e., the columns $\mathbf{u}_k \in \mathbb{R}^m$, $k = 1, \ldots, r$ of the matrix U, give inter-annual variations of the relative magnitude the k-th seasonal pattern $\mathbf{v}_k$. For instance, if the magnitude of the entries of the vector $\sigma_k \mathbf{u}_k$ decays with their index, then the magnitude of the corresponding seasonal pattern $\mathbf{v}_k$ decays with years.

A rank-p approximation of the matrix $A \in \mathbb{R}^{m \times n}$ is the matrix $A_p \in \mathbb{R}^{m \times n}$, such that $\mathrm{rank}(A_p) = p$. The rank-p SVD approximation of A is defined as

$$U\Sigma_p V^T = \sum_{k=1}^{p} \sigma_k \mathbf{u}_k \mathbf{v}_k^T, \tag{3.12}$$

where $A = U\Sigma V^T$ and $1 \leq p \leq r = \mathrm{rank}(A)$. The rank-$p$ SVD provides the best rank-p approximation of A in Frobenius norm [16].

The relative error of the rank-p SVD approximation is related to the singular values of A as follows:

$$\begin{aligned} \rho_p^{\mathrm{SVD}} &= \frac{\|A - U\Sigma_p V^T\|_F^2}{\|A\|_F^2} = \frac{1}{\|A\|_F^2} \|U\Sigma V^T - U\Sigma_p V^T\|_F^2 = \frac{1}{\|A\|_F^2} \|U(\Sigma - \Sigma_p)V^T\|_F^2 \\ &= \frac{1}{\|A\|_F^2} \|\Sigma - \Sigma_p\|_F^2 = \frac{\sum_{k=p+1}^{r} \sigma_k^2}{\sum_{k=1}^{r} \sigma_k^2}. \end{aligned} \tag{3.13}$$

[16].

The SVD algorithm was programmed in Python. First, we created the matrix by initializing the data matrix using `np.array`. Then, we used `numpy.linalg.svd()` to execute the SVD algorithm we applied to the respective data matrix. This function returns the matrices U, Σ, V^T.

3.4. Non-Negative Matrix Factorization

Similar to SVD, the Non-negative Matrix Factorization (NMF) obtains a (possibly) low-rank representation of a data matrix that captures dominant patterns, while reducing noise and dimensionality. Furthermore, NMF requires non-negativity of entries of both the data matrix and the two matrices into which the data matrix is factorized. The following points reveal why NMF can be applied to the data matrices discussed in this research:

- **Precipitation:** daily totals are nonnegative by definition, so $A \geq 0$ holds directly.

- **Hydrothermal:** if defined as a nonnegative function of temperature and precipitation (e.g., product of nonnegative components or an index with truncation), the resulting matrix is nonnegative; specify the exact construction.
- **Temperature:** from figure 3.1, we see that the entries in the integrated temperatures matrix are non-negative. Hence, NMF can be used on the integrated temperature matrix.

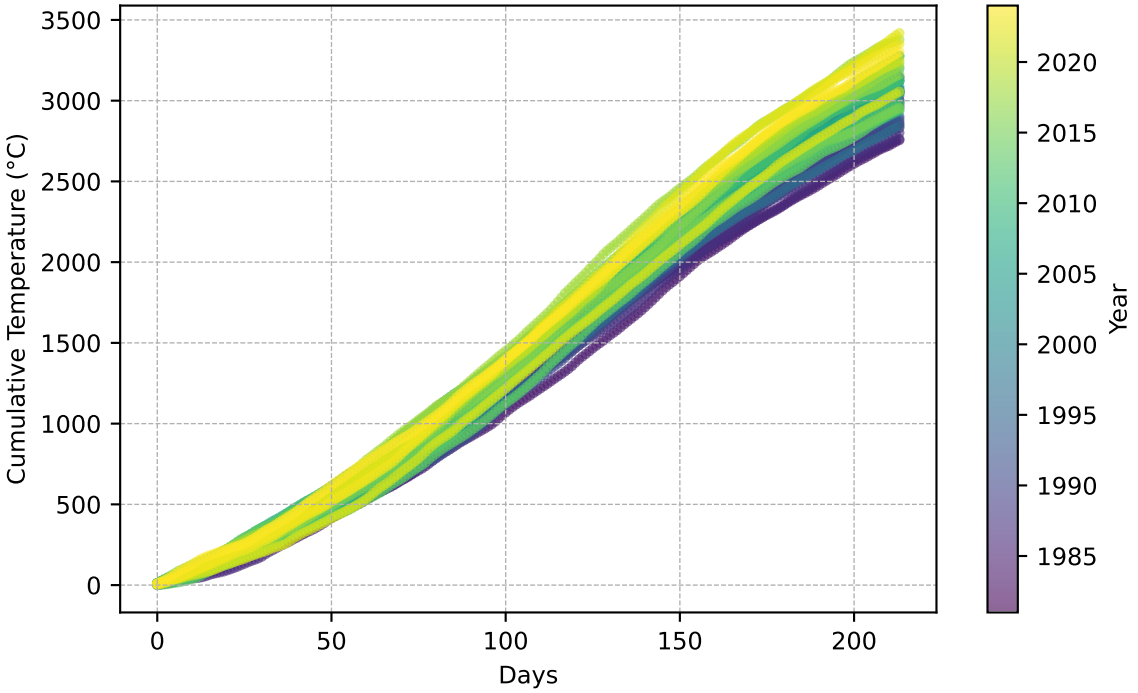

Figure 3.1: Illustration of the nonnegative integrated temperature matrix. All entries are ≥ 0.

Definition: Given a nonnegative matrix $A \in \mathbb{R}^{m \times n}_{\geq 0}$, rank-$p$ NMF finds nonnegative factors $W_p \in \mathbb{R}^{m \times p}_{\geq 0}$ and $H_p \in \mathbb{R}^{p \times n}_{\geq 0}$ that solve the minimization problem [17, 18]:

$$\min_{W_p, H_p \geq 0} \left\| A - W_p H_p \right\|_F^2.$$

Existence: Existence of NMF was showed for the first time by using the Completely Positive (CP) Factorization in [19].

Uniqueness: NMF is not unique. This means that matrix $A \approx W_p H_p$ can consist out of different matrices W_p and H_p, all depending on how many components were used when running the algorithm. The number of components we choose to work with when running the NMF algorithm directly affects the factorization space. A higher number of components gives us more degrees of freedom. This increases the number of possible factorizations [17].

Interpretation of the NMF matrices: When running the algorithm we expect the outcome to be as follows: $A \approx W_p H_p$. The matrices W_p and H_p are respectively the coefficient matrix and the basis matrix. Similar to SVD, the coefficients tell us how strong a pattern was on a yearly basis. While the basis vectors (rows of the matrix H_p) describe the patterns within the season.

Furthermore, when comparing NMF representations of different ranks, it is convenient to use the following relative error [20]:

$$\rho_p^{\text{NMF}} = \frac{\|A - W_p H_p\|_F^2}{\|A\|_F^2}. \tag{3.14}$$

(3.15)

Implementation and software: We have used an implementation of the NMF algorithm offered by the `sklearn.decomposition` Python package, namely, the function `NMF`. Application of the NMF algorithm to our data often produced convergence warnings. These were mostly due to the maximum number of iterations being set too low. This warning then signifies that the algorithm has simply failed to reach the required tolerance within this number of iterations. This can be prevented by increasing the maximum iterations. However, this was not the only reason for non-convergent behavior. Sometimes, the lack of convergence also had to do with the initialization that was used. The random initialization worked better than the SVD-based initialization. This has to do with the fact that environmental variables evolve in nonlinear, chaotic ways, while the SVD-based initialization is geared towards linear low-rank structures, which may not reflect the true dynamics. Also, random initializations explore more than one starting point [21].

3.5. Verifying generalizability of seasonal patterns

The seasonal patters, i.e., matrices V and H, obtained from the training datasets may be appropriate for other years as well. To check if this is the case, we consider the approximation of testing datasets A_{tst}, using V and H constructed by SVD and NMF on the training years.

Let $\mathbf{a}_t^T$ be a single data vector corresponding to some test year. Then, the problem of approximating the test year data using the seasonal patterns detected by the SVD in the training years can be formulated as the following least-squares minimization problem:

$$\min_{\mathbf{b}_t \in \mathbb{R}^p} \|V_p \mathbf{b}_t - \mathbf{a}_t\|_2^2, \tag{3.16}$$

where $V_p \in \mathbb{R}^{n \times p}$ is the matrix with p seasonal patterns as columns, i.e., the SVD matrix of right singular vectors. The solution of (3.16) is given by:

$$\mathbf{b}_{t,p} = V_p^T \mathbf{a}_t. \tag{3.17}$$

The error of this rank-p approximation is defined as:

$$\rho_{t,p}^{\text{SVD}} = \frac{\|\mathbf{a}_t - \mathbf{a}_{t,p}\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|\mathbf{a}_t - V_p \hat{\mathbf{b}}_t\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|\mathbf{a}_t - V_p V_p^T \mathbf{a}_t\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|(I - V_p V_p^T)\mathbf{a}_t\|_2^2}{\|\mathbf{a}_t\|_2^2} \tag{3.18}$$

[16].

In the NMF case, the matrix of seasonal patterns H is not an orthogonal matrix. Therefore the corresponding least-squares problem

$$\min_{\mathbf{b}_t \in \mathbb{R}^p} \|H_p^T \mathbf{b}_t - \mathbf{a}_t\|_2^2, \tag{3.19}$$

has a different solution:

$$\mathbf{b}_{t,p} = (H_p H_p^T)^{-1} H_p \mathbf{a}_t. \tag{3.20}$$

The error of this rank-p NMF-derived approximation is defined as:

$$\rho_{t,p}^{\text{NMF}} = \frac{\|\mathbf{a}_t - \mathbf{a}_{t,p}\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|\mathbf{a}_t - H_p^T \mathbf{b}_{t,p}\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|\mathbf{a}_t - H_p^T (H_p H_p^T)^{-1} H_p \mathbf{a}_t\|_2^2}{\|\mathbf{a}_t\|_2^2} = \frac{\|(I - H_p^T (H_p H_p^T)^{-1} H_p)\mathbf{a}_t\|_2^2}{\|\mathbf{a}_t\|_2^2} \tag{3.21}$$

[17].

3.6. Predicting the time evolution of seasonal patterns

To predict how seasonal patterns change over time, the temporal evolution of the component score vector (coefficients) of each pattern is modelled using a simple linear-in-time model. The parameters of this linear time-evolution model are estimated on the training data set of past years. The predictions of this model are evaluated on the test set of future years. In this section, we discuss the linear-in-time

regression analysis of the SVD and NMF coefficients describing interannual variations of seasonal EV patterns, employing metrics that characterize the usefulness and quality of linear regression. In particular, we discuss the slope, the p-value, and the coefficient of determination.

For each year t, the EV data vector $\mathbf{a}_t \in \mathbb{R}^n$ is projected onto a p-dimensional reduced basis $V_p = [\mathbf{v}_1 \cdots \mathbf{v}_p] \in \mathbb{R}^{n\times p}$ producing the approximation:

$$\mathbf{a}_{t,p} = \sum_{i=1}^{p} b_{t,i}\mathbf{v}_i = V_p\mathbf{b}_t \tag{3.22}$$

Here, $\mathbf{b}_t \in \mathbb{R}^p$ is the vector of p coefficients for the year t. Note that the seasonal patterns in V_p have been deduced from the EV data of the training years. Thus, the assumption is that the patterns themselves do not change over the years years, however, the relative weight of each pattern changes from year to year.

The change of the entries $b_{t,i}$ of $\mathbf{b_t}$ with time t will be modelled as a linear trend plus noise:

$$b_{t,i} = \alpha_i + \beta_i t + \epsilon_{t,i} \tag{3.23}$$

Here, t is the year, α_i is the intercept, β_i is the slope and $\epsilon_{t,i}$ refers to the residual error.

Parameters α_i and β_i used in equation (3.23) are estimated on the training years by using the Ordinary Least Squares (OLS) method. Let the training set of years be denoted by T_{tr}, then the OLS estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ minimize the following functional:

$$\sum_{t\in T_{\mathrm{tr}}} (b_{t,i} - (\alpha + \beta t))^2 \tag{3.24}$$

For each fitted component we report the slope $\hat{\beta}_i$, the p-value for testing $\beta_i = 0$, and the coefficient of determination R_i^2. This was done using the `SciPy` Python package, specifically the `scipy.stats` module.

The estimated parameters $\hat{\alpha}_i$ and $\hat{\beta}_i$ are obtained from the training years. These parameters are used to predict the temporal evolution of $b_{t,i}$ over the test years. For a test year t_{tst}, the predicted vector of coefficients is

$$\hat{\mathbf{b}}_{t_{\mathrm{tst}}} = \begin{bmatrix} \hat{\alpha}_1 + \hat{\beta}_1 t_{\mathrm{tst}} \\ \vdots \\ \hat{\alpha}_p + \hat{\beta}_p t_{\mathrm{tst}} \end{bmatrix}. \tag{3.25}$$

The predicted seasonal EV data vector is obtained with the basis vectors learned over the training years with the predicted coefficients:

$$\hat{\mathbf{a}}_{t_{\mathrm{tst}},p} = V_p\hat{\mathbf{b}}_{t_{\mathrm{tst}}}. \tag{3.26}$$

To assess the quality of these predictions, we compare the predicted and the obeserved rank-p representations of EV data-vectors. Prediction accuracy of the rank-p approximation is quantified using the Normalized Mean Squared Errors (NMSE's). Specifically, the rank-p prediction NMSE:

$$\mathrm{NMSE}_{\mathrm{pred}}(t_{\mathrm{tst}}) = \frac{\|\mathbf{a}_{t_{\mathrm{tst}}} - \hat{\mathbf{a}}_{t_{\mathrm{tst}},p}\|_2^2}{\|\mathbf{a}_{t_{\mathrm{tst}}}\|_2^2}, \tag{3.27}$$

and the rank-p approximation NMSE:

$$\mathrm{NMSE}_{\mathrm{app}}(t_{\mathrm{tst}}) = \frac{\|\mathbf{a}_{t_{\mathrm{tst}}} - \mathbf{a}_{t_{\mathrm{tst}},p}\|_2^2}{\|\mathbf{a}_{t_{\mathrm{tst}}}\|_2^2}, \tag{3.28}$$

which corresponds to the previously defined errors (3.18) and (3.21) [22].

Slope: The slope $\hat{\beta}_i$ is derived by minimizing the LS functional defined in equation (3.24):

$$\begin{aligned}\sum_{t\in T_{tst}}(b_{t,i}-(\alpha+\beta t))^2 &= \sum_{j=1}^{n}(b_{t_j,i}-(\alpha+\beta t_j))^2 \\ &= \sum_{j=1}^{n}(b_{t_j,i}-\begin{bmatrix}1 & t_j\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix})^2 = \|\boldsymbol{y}_i - X\boldsymbol{\alpha}\|_2^2\end{aligned} \tag{3.29}$$

Here $t_j \in T_{\text{tr}} = [t_1, \ldots, t_n]$ are the training years. Furthermore,

$$X = \begin{bmatrix}1 & t_1\\ 1 & t_2\\ \vdots & \vdots\\ 1 & t_n\end{bmatrix} \in \mathbb{R}^{n\times 2}, \quad \boldsymbol{\alpha} = \begin{bmatrix}\alpha\\ \beta\end{bmatrix} \in \mathbb{R}^2, \quad \boldsymbol{y_i} = \begin{bmatrix}b_{t_1,i}\\ b_{t_2,i}\\ \vdots\\ b_{t_n,i}\end{bmatrix} \in \mathbb{R}^n \tag{3.30}$$

The ordinary least squares (OLS) solution satisfies the following normal equation:

$$X^T X\boldsymbol{\alpha} = X^T\boldsymbol{y}_i, \tag{3.31}$$

and is obtained as:

$$\hat{\boldsymbol{\alpha}}_i = (X^T X)^{-1}X^T\boldsymbol{y}_i. \tag{3.32}$$

First, we write out the result for $X^T X$ as follows:

$$X^T X = \begin{bmatrix}1 & \cdots & 1\\ t_1 & \cdots & t_n\end{bmatrix}\begin{bmatrix}1 & t_1\\ \vdots & \vdots\\ 1 & t_n\end{bmatrix} = \begin{bmatrix}n & \sum_{j=1}^n t_j\\ \sum_{j=1}^n t_j & \sum_{j=1}^n (t_j)^2\end{bmatrix} \tag{3.33}$$

Now, we write out the result for $X^T\boldsymbol{y}_i$ as follows:

$$X^T\boldsymbol{y}_i = \begin{bmatrix}n & \cdots & 1\\ t_1 & \cdots & t_n\end{bmatrix}\begin{bmatrix}b_{t_1,i}\\ \vdots\\ b_{t_n,i}\end{bmatrix} = \begin{bmatrix}\sum_{j=1}^n t_j b_{t_j,i}\\ \sum_{j=1}^n b_{t_j,i}\end{bmatrix} \tag{3.34}$$

Now substituting equation (3.33) and equation (3.34) in equation (3.32) gives the following result:

$$\begin{aligned}\begin{bmatrix}\hat{\alpha}_i\\ \hat{\beta}_i\end{bmatrix} &= \begin{bmatrix}n & \sum t_j\\ \sum t_j & \sum t_j^2\end{bmatrix}^{-1}\begin{bmatrix}\sum b_{j,i}\\ \sum t_j b_{j,i}\end{bmatrix} \\ &= \frac{1}{n\sum t_j^2 - (\sum t_j)^2}\begin{bmatrix}\sum t_j^2 & -\sum t_j\\ -\sum t_j & n\end{bmatrix}\begin{bmatrix}\sum b_j\\ \sum t_j b_j\end{bmatrix} \\ &= \frac{1}{n\sum t_j^2 - (\sum t_j)^2}\begin{bmatrix}\sum t_j^2\sum b_j - \sum t_j\sum t_j b_j\\ -\sum t_j\sum b_j + n\sum t_j b_j\end{bmatrix}\end{aligned} \tag{3.35}$$

From equation (3.35), we can derive the equations for the intercept and slope, respectively, as [23]:

$$\hat{\alpha}_i = \frac{\sum t_j^2\sum b_j - \sum t_j\sum t_j b_j}{n\sum t_j^2 - (\sum t_j)^2} \tag{3.36}$$

$$\hat{\beta}_i = \frac{-\sum t_j\sum b_j + n\sum t_j b_j}{n\sum t_j^2 - (\sum t_j)^2} \tag{3.37}$$

p-value: The p-value quantifies the probability of observing a slope estimate at least as extreme as $\hat{\beta}_i$ under the null hypothesis $H_0 : \beta_i = 0$. In practice, we compare the p-value to a chosen significance level. In this study, we opted for $\alpha = 0.05$. If $p < 0.05$, we reject H_0 and conclude that the slope β_i is significantly different from zero. This suggests that there is evidence of a linear trend in $b_{t,i}$ over time.

If $p \geq 0.05$, we do not reject H_0. In this case, the data do not provide sufficient evidence to claim that β_i differs from zero, and the apparent trend may be due to random variation.

The test statistic is given by

$$t = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)}, \tag{3.38}$$

where $\mathrm{SE}(\hat{\beta}_i)$ denotes the standard error of the slope estimate. The standard error is obtained from the estimated variance of the residuals,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^{n} \left(b_{t_j,i} - (\hat{\alpha}_i + \hat{\beta}_i t_j)\right)^2, \tag{3.39}$$

and the variance of $\hat{\beta}_i$ is

$$\mathrm{Var}(\hat{\beta}_i) = \frac{\hat{\sigma}^2}{\sum_{j=1}^{n}(t_j - \bar{t})^2}. \tag{3.40}$$

Thus,

$$\mathrm{SE}(\hat{\beta}_i) = \sqrt{\mathrm{Var}(\hat{\beta}_i)}. \tag{3.41}$$

The p-value is then computed as

$$p = 2 \cdot \left(1 - F_{t_{n-2}}(|t|)\right), \tag{3.42}$$

where $F_{t_{n-2}}$ is the cumulative distribution function of the Student's t distribution with $n-2$ degrees of freedom [23].

Coefficient of determination: The coefficient of determination quantifies how well the linear regression model explains the variability of the data. It is defined as follows:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}, \tag{3.43}$$

Here, SSE is the residual sum of squares and SST is the total sum of squares. The range of R^2 values is $0 \leq R^2 \leq 1$,. $R^2 = 0$ indicates that the model does not explain any of the variability in the data; predictions are no better than using the mean. Meanwhile, $R^2 = 1$ indicates that the model explains all variability perfectly; the fitted line passes through all data points. Intermediate values $0 < R^2 < 1$ indicate that the model explains part of the variability. Larger values indicate a better fit [23].

Let $\bar{b}_i = \frac{1}{n}\sum_{j=1}^{n} b_{t_j,i}$. Define

$$\mathrm{SST} = \sum_{j=1}^{n}(b_{t_j,i} - \bar{b}_i)^2, \quad \mathrm{SSE} = \sum_{j=1}^{n}(b_{t_j,i} - \hat{\alpha}_i - \hat{\beta}_i t_j)^2, \tag{3.44}$$

and $\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$. The coefficient of determination is

$$R_i^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{j=1}^{n}(b_{t_j,i} - \hat{\alpha}_i - \hat{\beta}_i t_j)^2}{\sum_{j=1}^{n}(b_{t_j,i} - \bar{b}_i)^2}. \tag{3.45}$$

Train-test split: We apply the common $80-20\%$-rule, using 80% of the data to train the regression model and 20% to test its predictions. This balances reliable parameter estimation with independent evaluation [24]. Since dataset sizes vary, the effective split differs. This impacts both the stability of training and the robustness of prediction accuracy.

4

Temperature

Temperature is a key environmental variable in this study, with direct implications for the agriculture. This chapter focuses on predicting long-term temporal trends. Insection 4.1, both the temperature matrix and its integrated form are visualized. The integrated representation emphasizes cumulative seasonal patterns rather than short-term fluctuations. This provides a more suitable basis for identifying dominant temporal trends across years. Subsequently, in section 4.2 and section 4.3, the results of SVD applied to the integrated temperature matrix are discussed. The components contributing most to the dominant structure of the matrix are selected for further predictive analysis. These predictions are assessed by comparing the NMSE values of both the rank-1 approximation and linear prediction. Then the same procedure is repeated using NMF in section 4.4. This enables a direct comparison between orthogonal and non-orthogonal decompositions, and provides insight into which representation offers the most robust characterization of the temperature data.

4.1. Exploratory Analysis of the Integrated Temperature Matrix

The entries of a few columns and rows of the temperature data matrix were introduced in section 2.2. In this section, both the temperature matrix and the integrated temperature matrix are visualized.

Figure 4.1 shows the first three rows of the temperature matrix. The daily temperature patterns across the years appear to be relatively similar. This can be confirmed by figure 4.2, where the integrated temperature values over the season only show minor deviations between 1981 and 1983. The integrated representation emphasizes cumulative seasonal patterns rather than short-term fluctuations. This is more suitable for identifying dominant patterns. Further analysis will be conducted on the integrated temperature matrix.

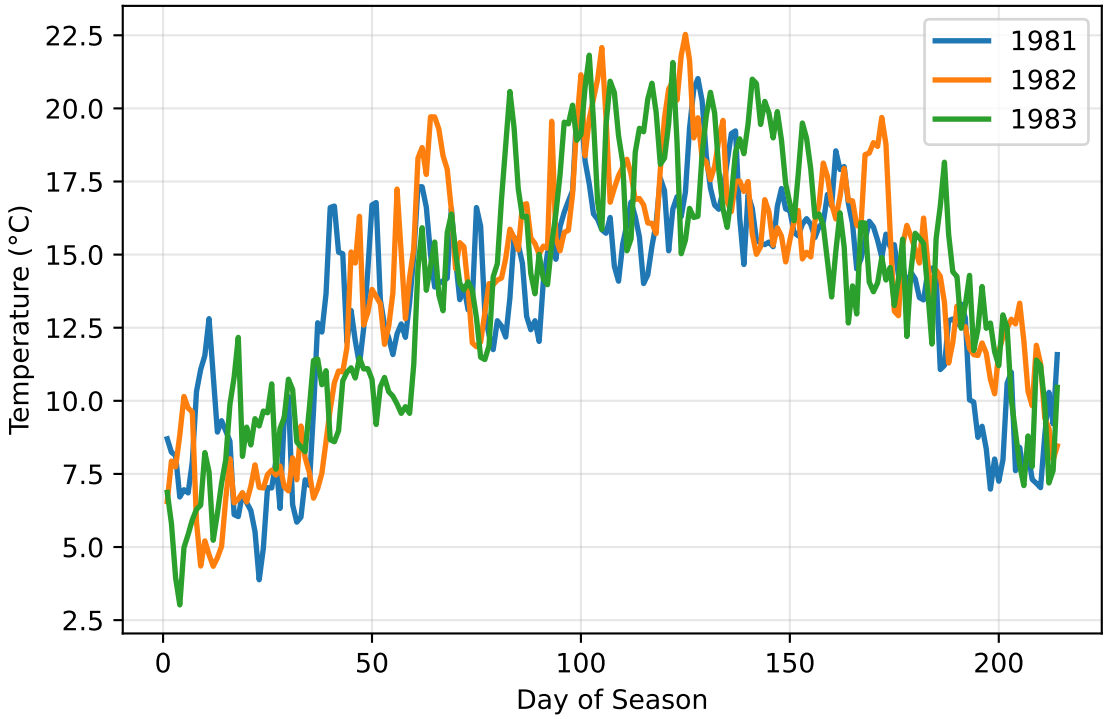

Figure 4.1: Temporal patterns across the season for the years 1981-1983. The many peaks and dips signify short-term fluctuations, rather than long-term temporal behavior.

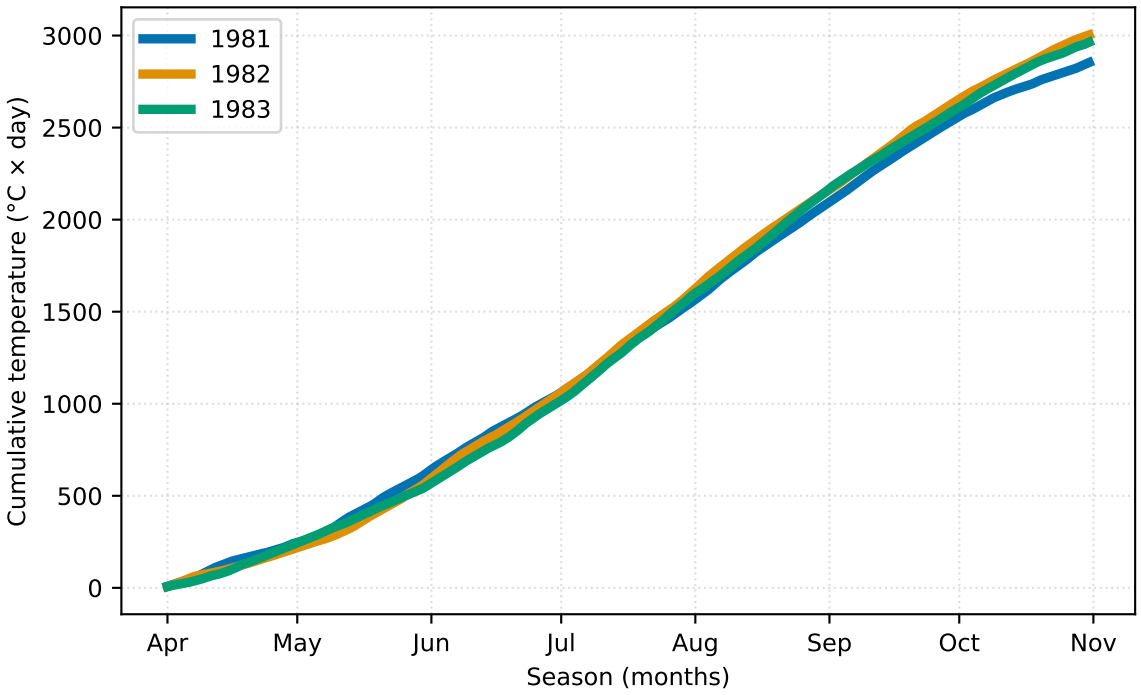

Figure 4.2: Cumulative temperature patterns across the years. These curves display long-term temporal behavior.

4.2. Characterizing Integrated Temporal Patterns through SVD

Singular Value Decomposition (SVD) was applied to the integrated temperature matrix. The analysis starts by analyzing which ranks capture most of the dominant structure of the integrated temperature matrix. Furthermore, the interpretation of these corresponding components is provided. The method used in this analysis can be found in section 3.3.

Figure 4.3 displays the relative approximation error as a function of the rank. The SVD error approximation ρ_p^{SVD} approaches 0.005% at rank=3. This means that the integrated temperature matrix is well approximated by a low-rank representation. This implies that the dominant structure of the matrix is captured by the first three basis vectors.

In figure 4.4, the seasonal temporal patterns are illustrated. The first basisvector highlights the slow cooling trend across the season. The second basisvector emphasizes an early summer followed by rapid cooling. This is reflective to an early start of autumn. The third basisvector contains two distinct peaks. These peaks mathematically satisfy the orthogonality constraints of SVD. However, they do not correspond to typical temporal behavior in the Netherlands. This is a clear pitfall of SVD: while it provides an optimal low-rank decomposition in a least-squares sense, the resulting basisvectors are not guaranteed to have direct physical meaning, due to the orthogonality constraint.

In figure 4.5, the interannual temporal coefficients are shown. The first coefficient remains negative across the years, emphasizing the pattern displayed by the first basis vector which took on negative values across the season. By contrast, the second and third temporal coefficients exhibit oscillations, which do not contribute meaningfully to the characterization of their associated basis vectors.

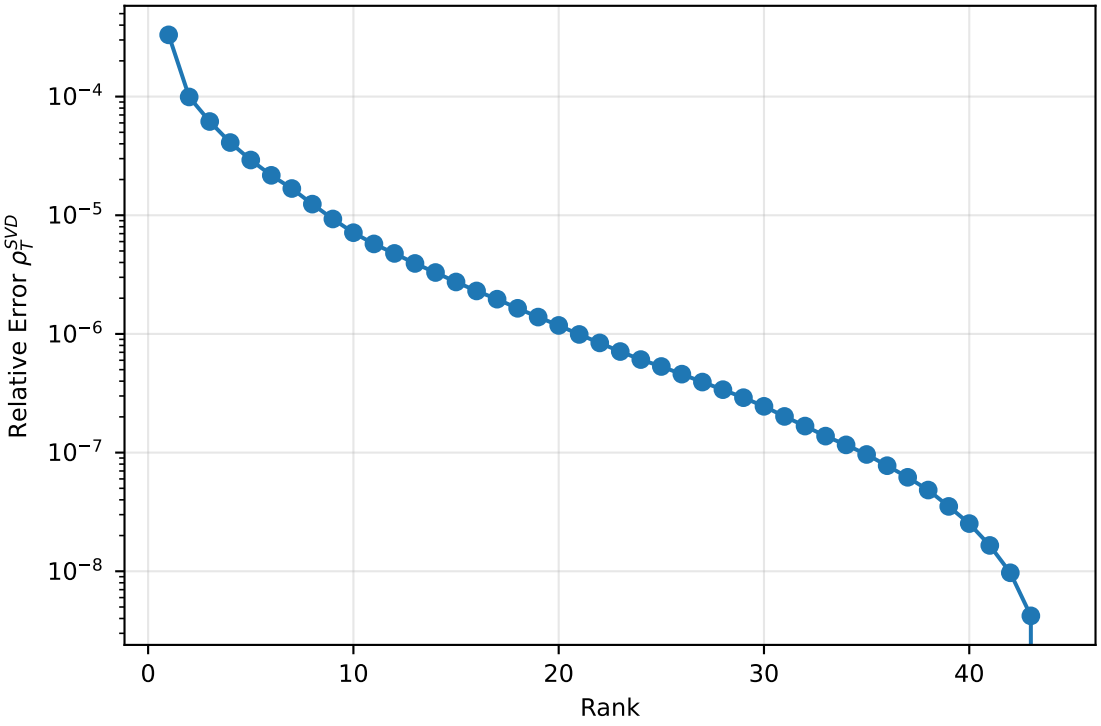

Figure 4.3: SVD approximation error ρ_p^{SVD} as a function of the approximation rank p for the integrated temperature matrix.

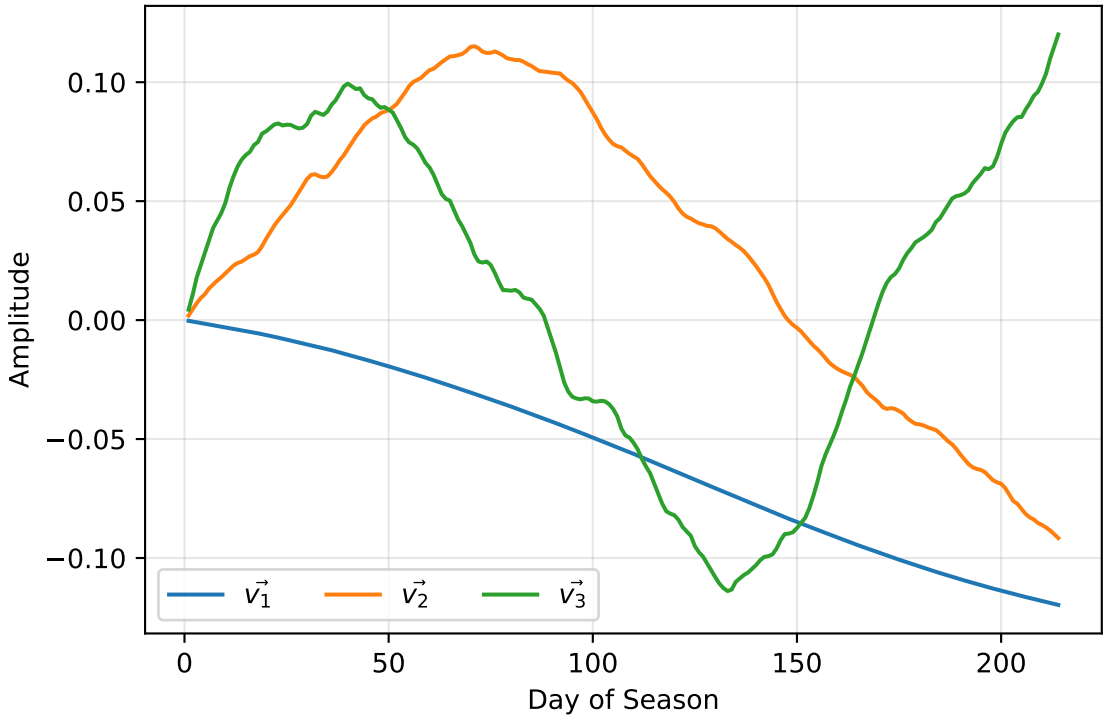

Figure 4.4: The three corresponding basis vectors obtained through SVD. These curves show the evolution of the temperature patterns across the season.

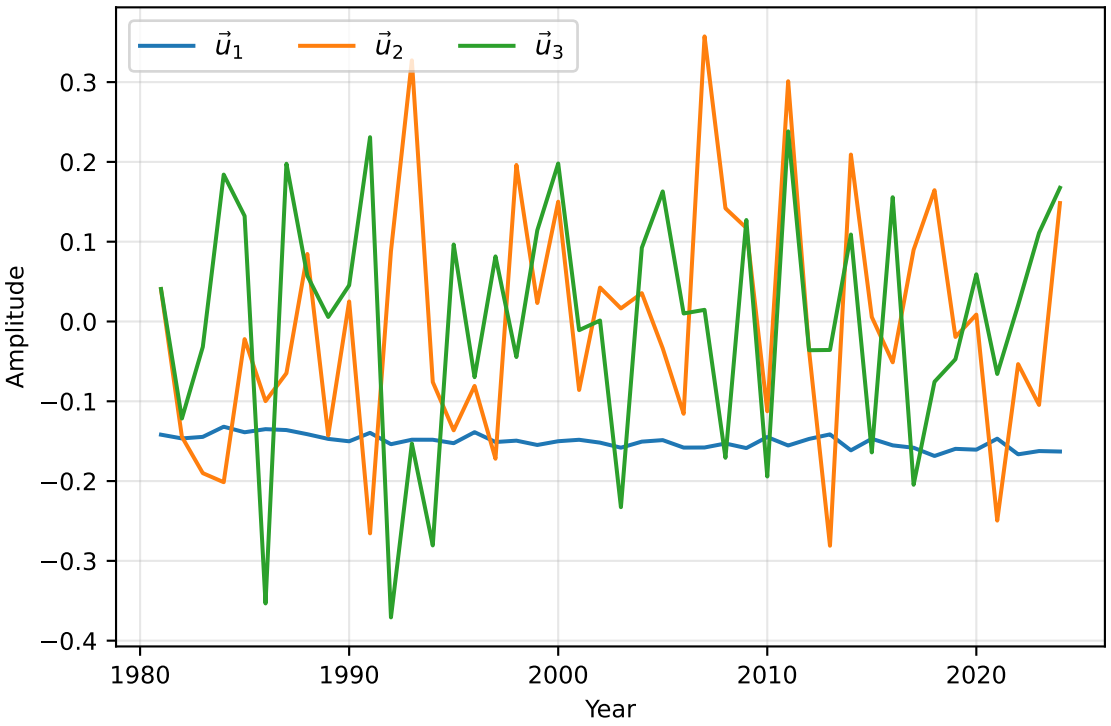

Figure 4.5: The coefficients corresponding to the seasonal temperatoral basis vectors generated by SVD.

4.3. Pattern Forecasting

In this section, the slopes, p-values, and coefficient of determination are examined of each component obtained through SVD. Based on these results, the relevant component(s) are selected for forecasting temporal patterns. The methods used in this section can be found in section 3.6.

In figure 4.6 the slopes and p-values of all the components obtained through SVD are displayed. The first slope has the largest absolute magnitude, emphasizing the importance of the trend depicted by the first basis vector. Its corresponding p-value is less than 0.05, implying statistical significance of the first basis vector. This basis vector also has the largest coefficient of determination as shown in figure 4.7. Although the basis vectors $7, 10, 31$ also have a p-value less than 0.05 they do not exhibit large slopes nor do they exhibit large coefficients of determination. Hence, only the first basis vector will be taken into account when predicting temporal patterns.

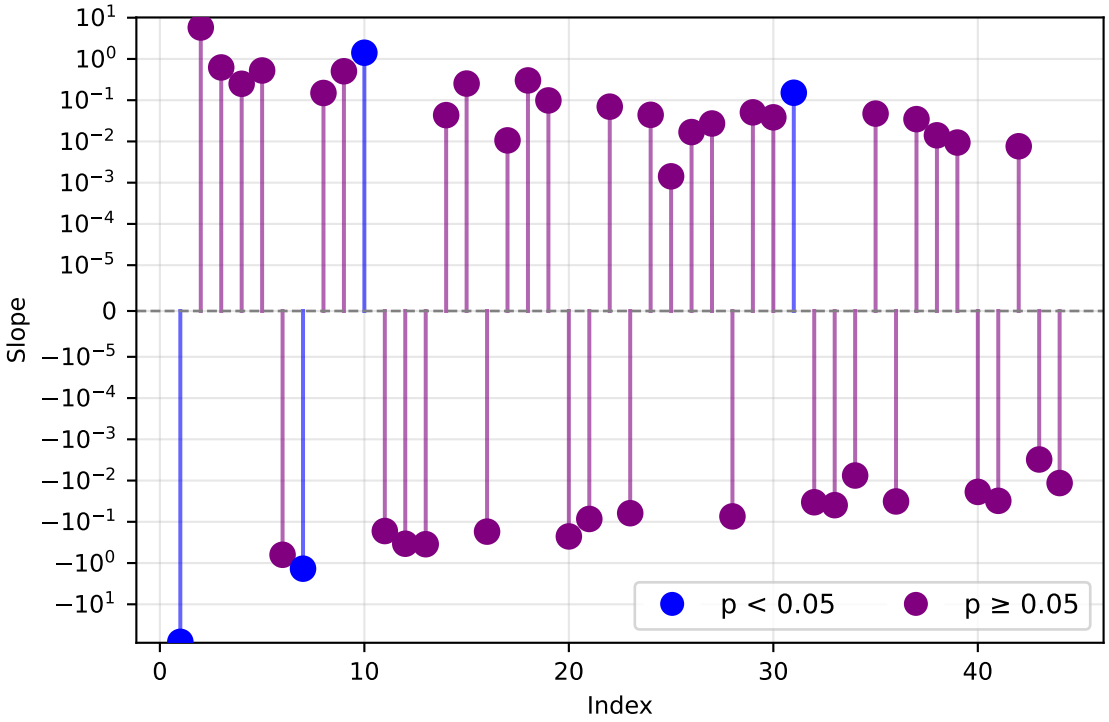

Figure 4.6: Slopes and associated p-values for the components obtained through SVD. The first component exhibits the largest absolute slope. Components $1, 7, 31$ exhibit p-value less than 0.05.

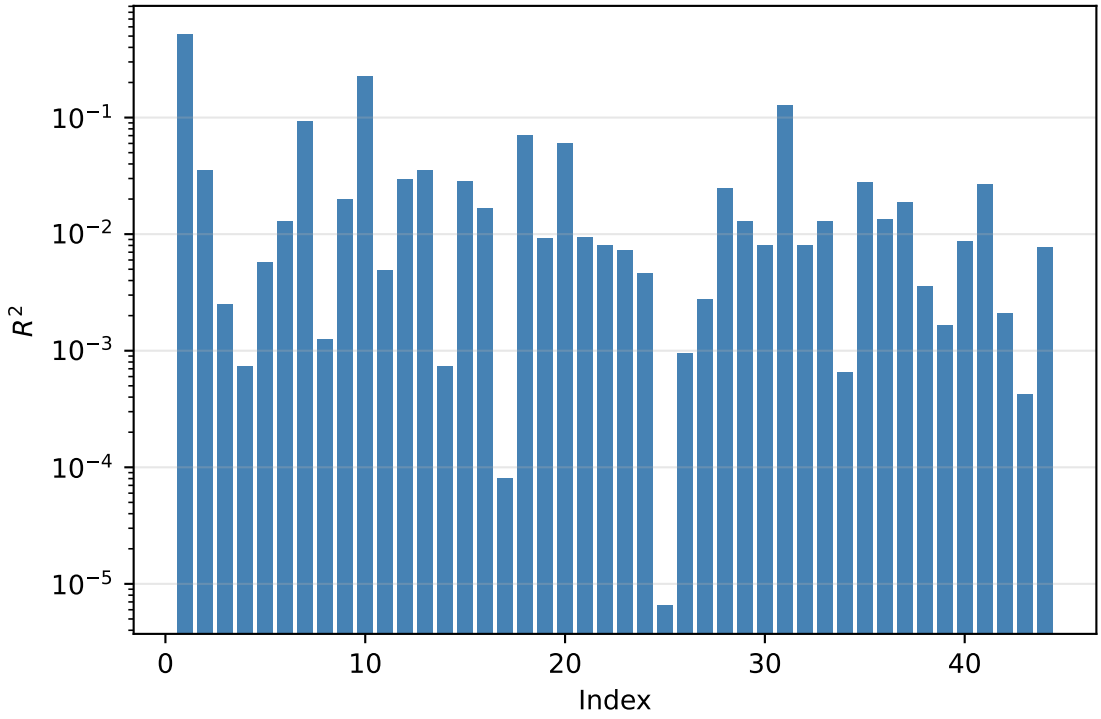

Figure 4.7: Coefficient of determination (R^2) for all components obtained through SVD. The first component has the highest coefficient,indicating the strongest explanatory power.

In figure 4.8, the NMSE values of both the rank-1 approximations $NMSE_{approx}(t_{tst})$, and linear prediction $NMSE_{pred}(t_{tst})$ are shown. $NMSE_{approx}(t_{tst})$ consistently yields lower NMSE values. This indicates that the dominant structure of the data is best captured by the first singular component rather than by a simple linear prediction. Between 2008 and 2012, both the $NMSE_{approx}(t_{tst})$, and $NMSE_{pred}(t_{tst})$ exhibit approximately the same NMSE values, due to a sudden shift in the temperature pattern [25]. The rank-1 approximation was still able to capture the data structure well during the rest of the test years, resulting in lower $NMSE_{approx}(t_{tst})$ values. By contrast, the linear prediction failed to represent the evolving trend of the temperature data. This suggests that this simple prediction could not adapt to the more complex temperature patterns in the later years. Finally, during 2017-2020, the NMSE values of both models dropped sharply again, indicating a return to more regular patterns in the data.

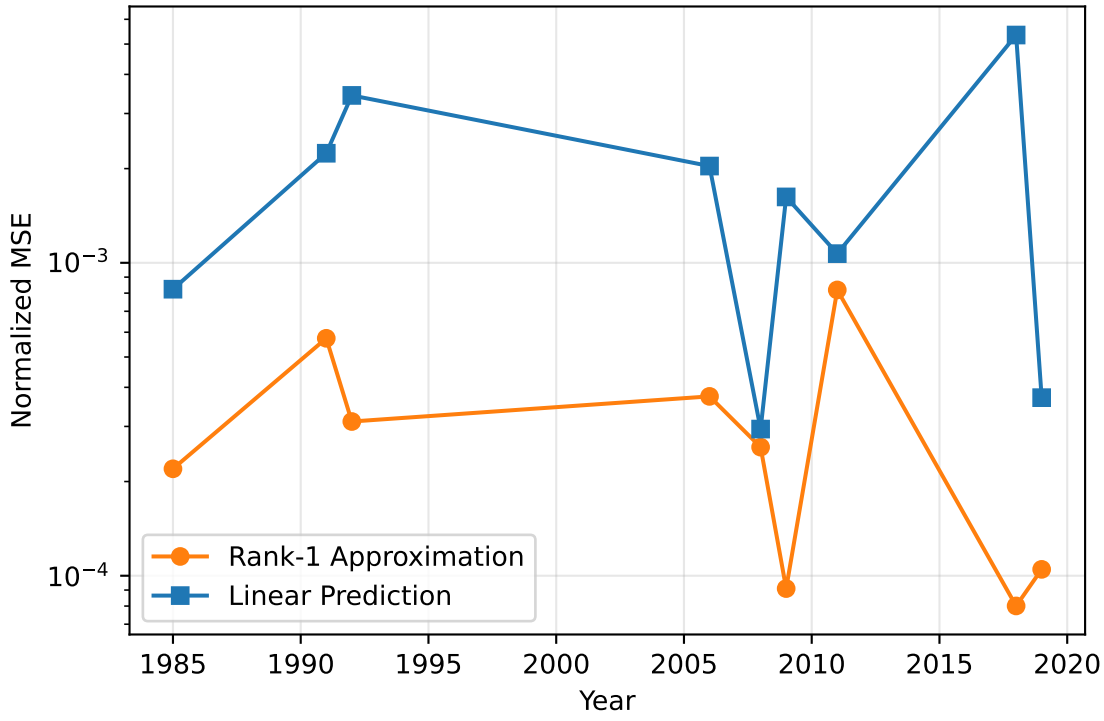

Figure 4.8: Comparison of NMSE values for the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$).

4.4. Integrated Temporal Trends Identified Through NMF

This section applies non-negative matrix factorization (NMF) to the integrated temperature matrix. The analysis considers convergence behavior across different ranks, followed by an interpretation of the resulting basis vectors and their coefficients. Finally, the rank-1 approximation is compared with a linear prediction to evaluate its ability to capture temporal patterns. The methods used in this analysis are discussed in section 3.4.

Figure 4.9 presents the relative approximation error (ρ_T^{NMF}) obtained for different ranks in the NMF algorithm. For the first three components, the algorithm converged within the maximum number of iterations, meanwhile for higher order components between 8 and 13 the algorithm failed to converge. This reflects the risk of the NMF getting stuck in suboptimal local minima when the factorization rank increases. This highlights the trade-off between model complexity and algorithmic stability: while higher ranks capture additional structure, they also increase the risk of non-convergence. Further analyses will be conducted using the three components, as the algorithm converged for these components, so the results are reliable.

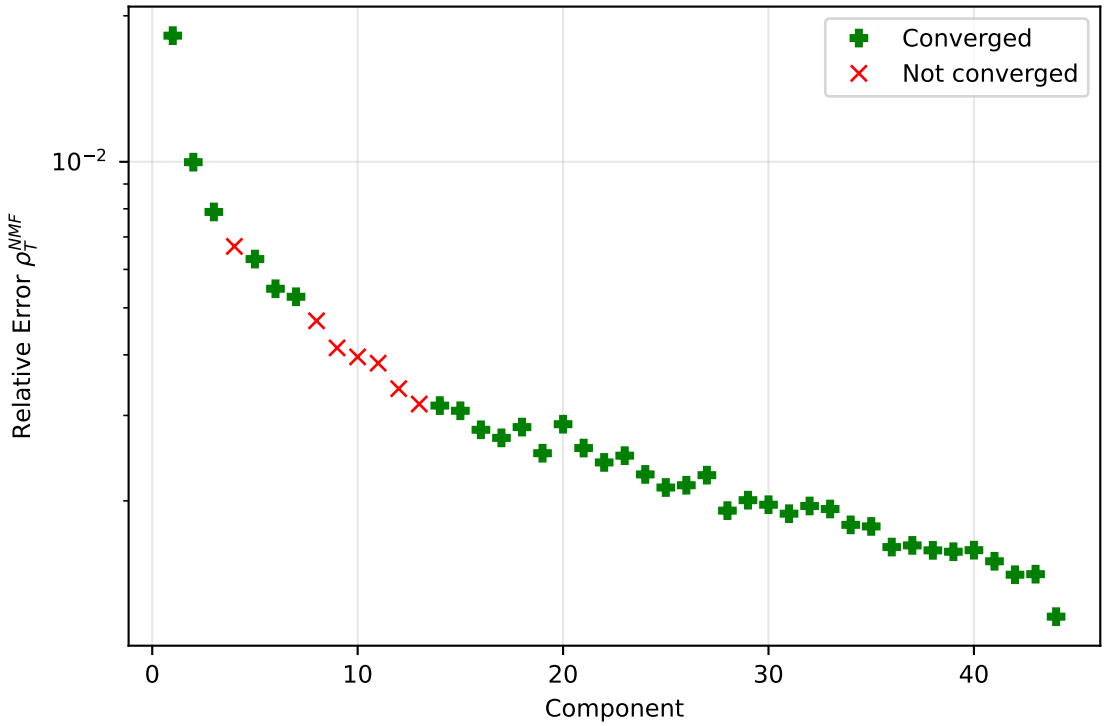

Figure 4.9: NMF approximation error ρ_T^{NMF} as a function of the approximation rank for the integrated temperature data matrix.

Figure 4.10, and figure 4.11 display the slopes, pvalues, and coefficient of determination for all NMF components. In general, components with higher indices exhibit smaller slopes and lack statistical significance, with the exception of components 23 and 26. Furthemore, we see that the first component has the largest slope and p-value less than 0.05. The second component also has a relatively large slope and p-value less than 0.05 with a higher coefficient of determination. Hence, both NMF obtained components are candidates for predicting temporal patterns.

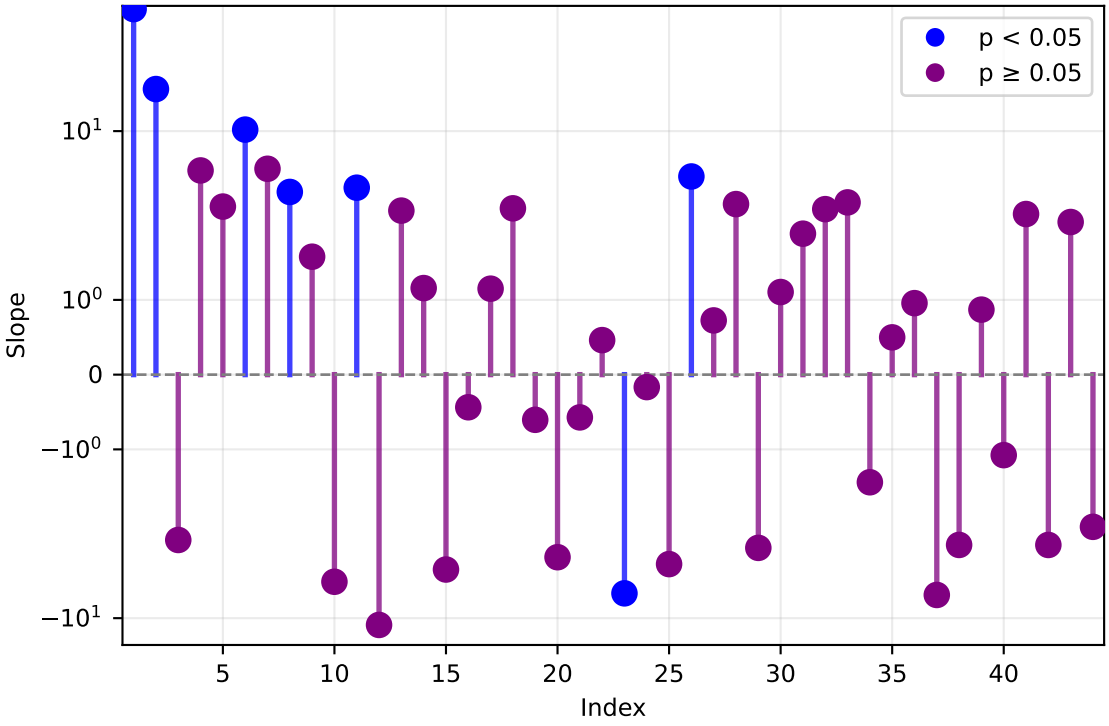

Figure 4.10: Slopes and p-values for all components obtained through NMF. The first and second components exhibit the largest slopes and p-values below 0.05, implying statistical significance.

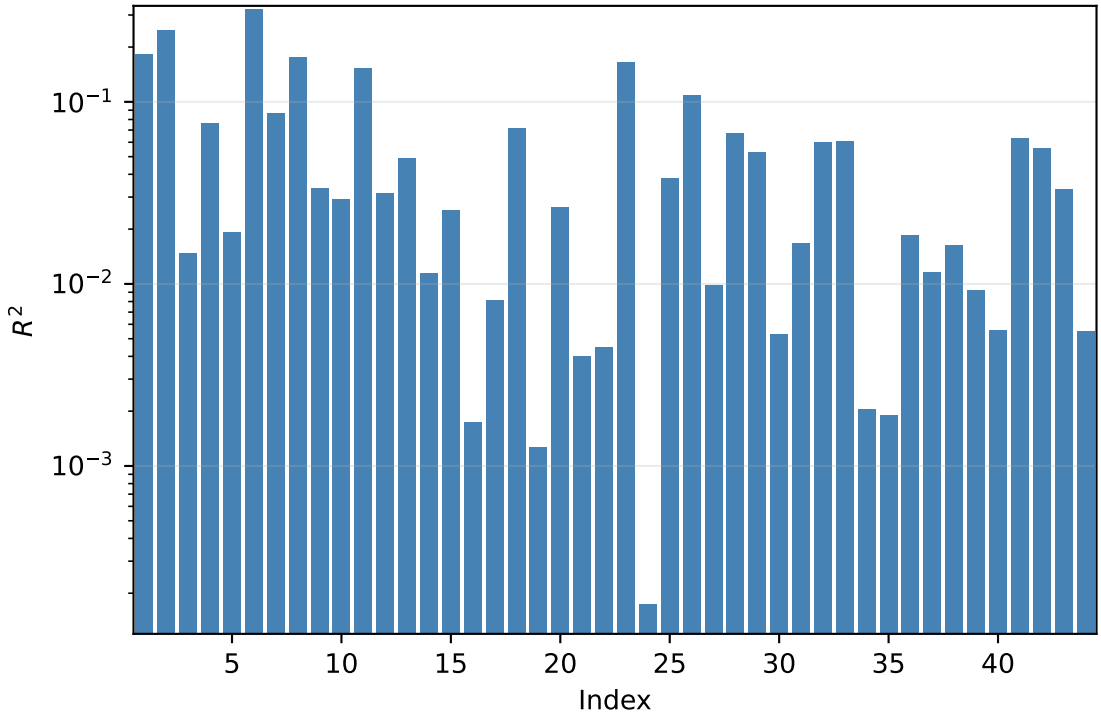

Figure 4.11: Coefficient of determination (R^2) for all components obtained through NMF. All components have R^2-value less than 10^{-2}. Furhtermore, the second component has a higher R^2-value compared to the first component.

The first two components appear to be suitable candidates for predicting temporal patterns. However,

unlike in SVD, the first component does not exhibit a markedly stronger contribution. This ambiguity arises because the basis vectors obtained through non-negative matrix factorization (NMF) are not orthogonal. Consequently, variance cannot be uniquely attributed to individual components.

To address this, the NMF algorithm is rerun using three components. Both the basis vectors and the corresponding coefficients depend on the number of components specified during the factorization process [17]. The choice of three components is motivated by the observation that the first two components exhibit the largest positive slopes and are statistically significant. The third component captures residual variation which is not explained by the previous two basis vectors. This ensures that the decomposition accounts not only for the dominant seasonal patterns but also for secondary fluctuations. The resulting basis vectors are presented in figure 4.12. The first and third basis vectors display similar seasonal patterns, while the second captures residual variations. The associated coefficients, shown in figure 4.13, remain positive across all years. This indicates that the patterns represented by the basis vectors are consistently relevant. The third coefficient obtains the smallest values overall, implying that the third basis vector is indeed a residual variation. Based on these results, the first component is selected for predicting temporal patterns.

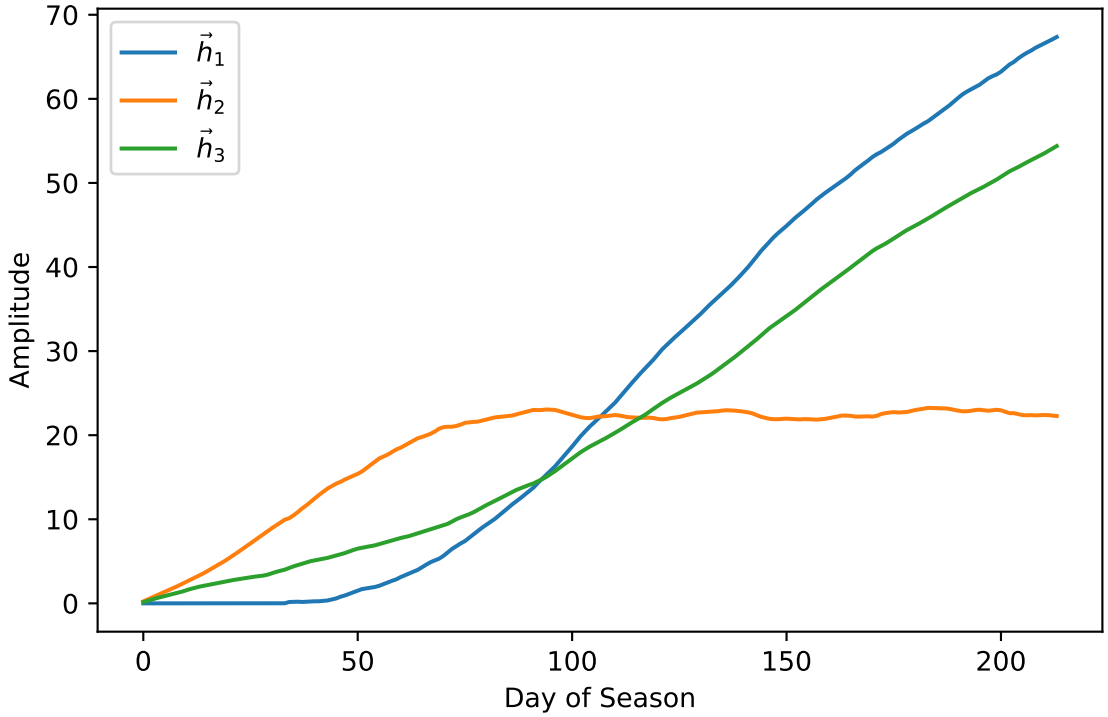

Figure 4.12: The three temporal basis vectors generated when rerunning the NMF algorithm using three components. These are the seasonal temporal patterns. The first and third basis vectors display similar behavior, where as the second basis vector displays residual variations.

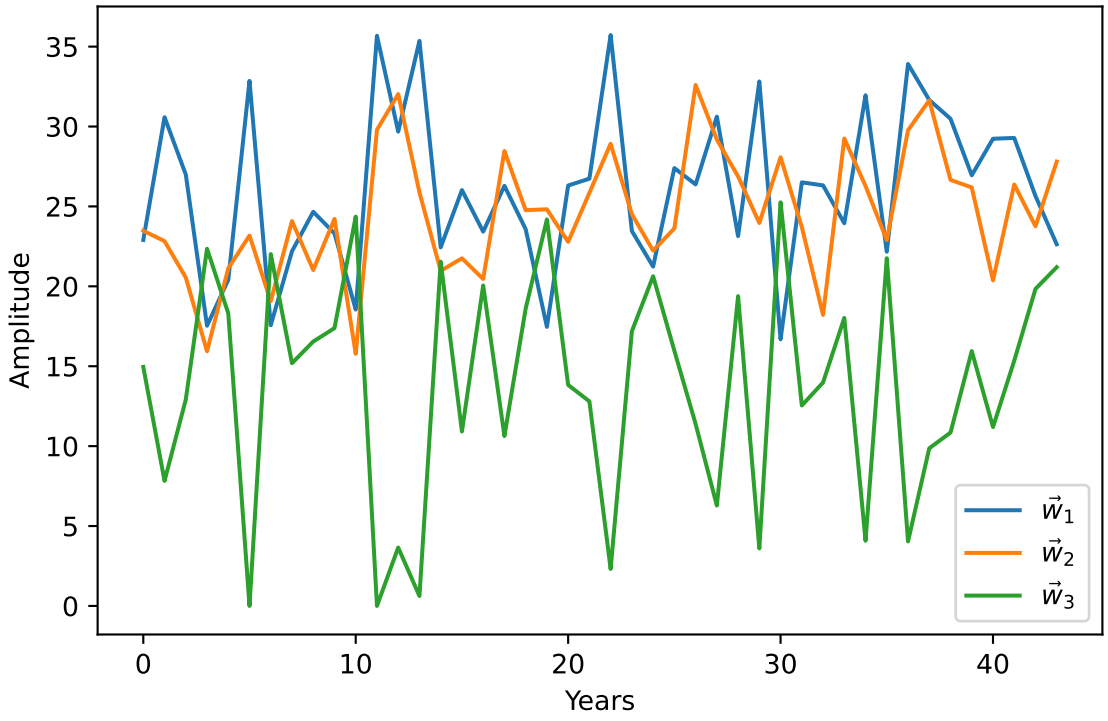

Figure 4.13: The temporal coefficients corresponding to the basis vectors obtained through NMF. The first coefficient obtains the highest values, highlighting the importance of the first basis vector.

In figure 4.14, the NMSE values of the rank-1 approximation ($NMSE_{approx}(t_{tst})$), and linear prediction ($NMSE_{pred}(t_{tst})$) are presented. These results are consistent with those shown in figure 4.8. However, the error peaks are smaller, which indicates that the approximations more accurately capture the structure of the basis vectors derived from NMF.

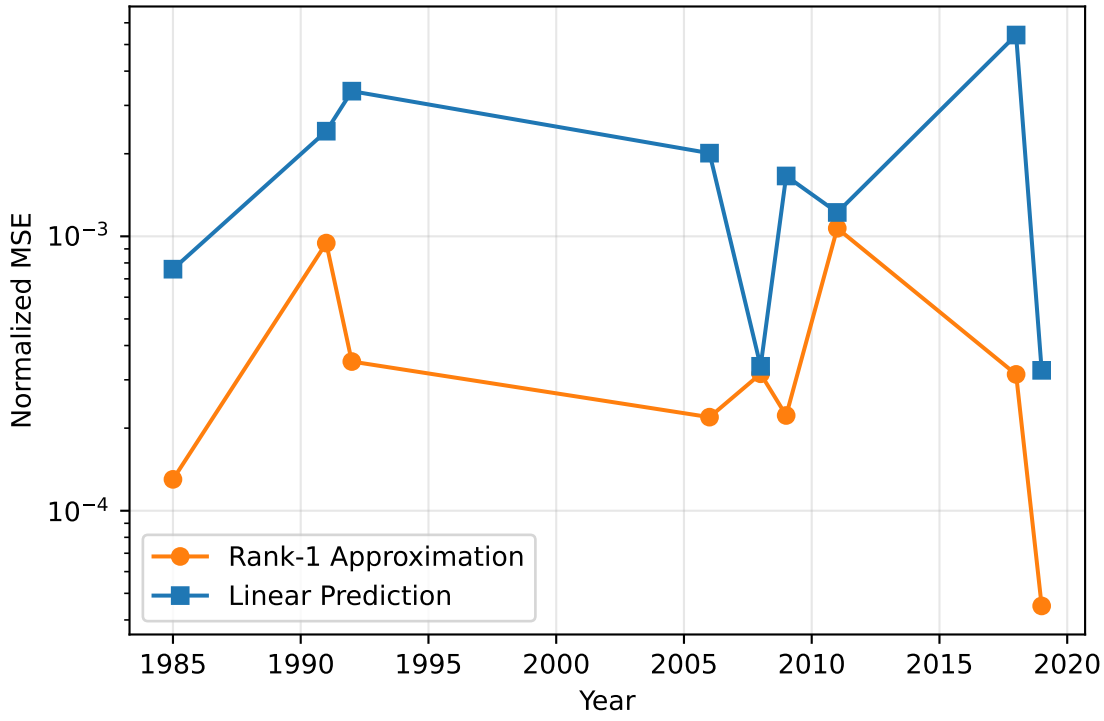

Figure 4.14: Comparison of NMSE values for the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$). The results are similar to the ones shown in figure 4.8.

5

Precipitation

Precipitation is another key environmental variable in this study. In recent years, notable shifts in precipitation patterns have been observed in the Netherlands. These changes have had significant consequences for crop yield and agricultural planning. This chapter examines long-term precipitation trends in the Netherlands by analyzing the integrated precipitation matrix, which emphasizes cumulative seasonal behavior rather than short-term variability in section 5.1. Singular value decomposition (SVD) and non-negative matrix factorization (NMF) are applied to identify dominant precipitation components and to evaluate their predictive value in section 5.2,and section 5.3. This provides the basis for assessing which representation offers the most robust characterization of precipitation dynamics in the Netherlands.

5.1. Visualization Of The Precipitation Matrix

In section 2.3, the preprocessing of the daily precipitation data and the construction of the precipitation matrix were described. In equation (2.2), the entries of a few columns of the precipitation matrix were introduced. In this section, the first three rows of the matrix and its integrated form are visualized. Figure 5.1 illustrates the precipitation patterns for the years 1930–1932. Based on the raw curves, it is difficult to determine which year was the wettest or driest, as all three contain several peaks across the season. Integration of the seasonal curves reveals that 1930 was the wettest year and 1931 the driest, as indicated by their final integrated values at the end of the season, as shown in figure 5.2.

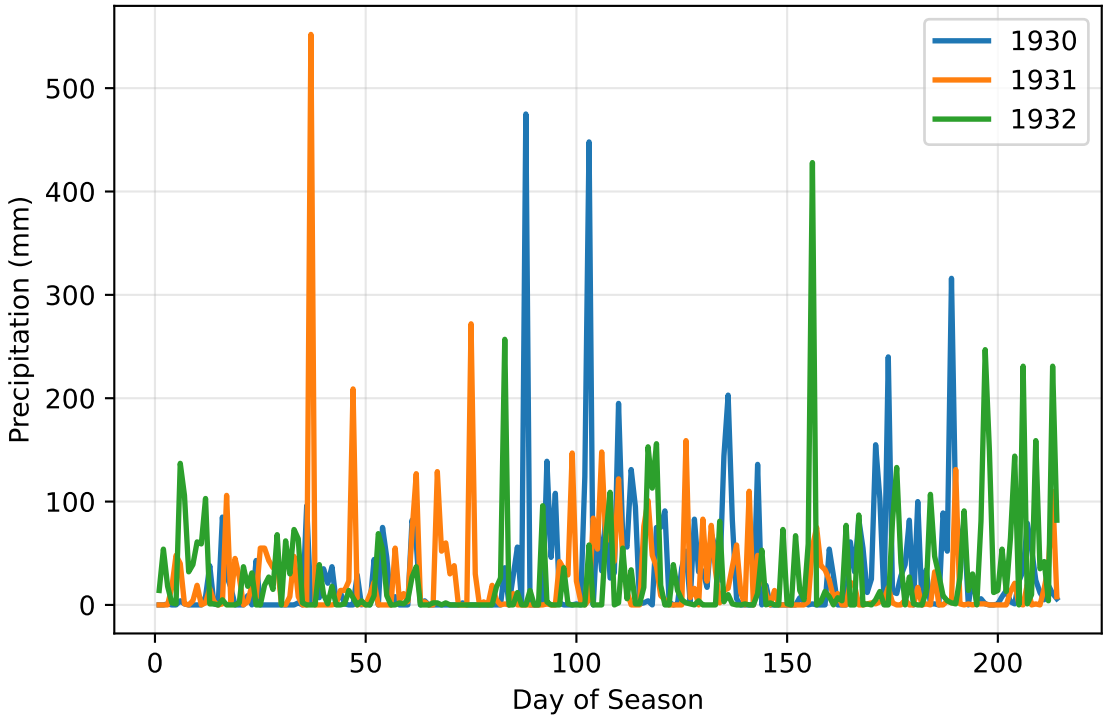

Figure 5.1: The seasonal precipitation patterns for the years between 1930 and 1932. These are the first three rows of the raw precipitation matrix. The peaks and dips indicate short-term fluctuations, rather than long-term precipitation behavior.

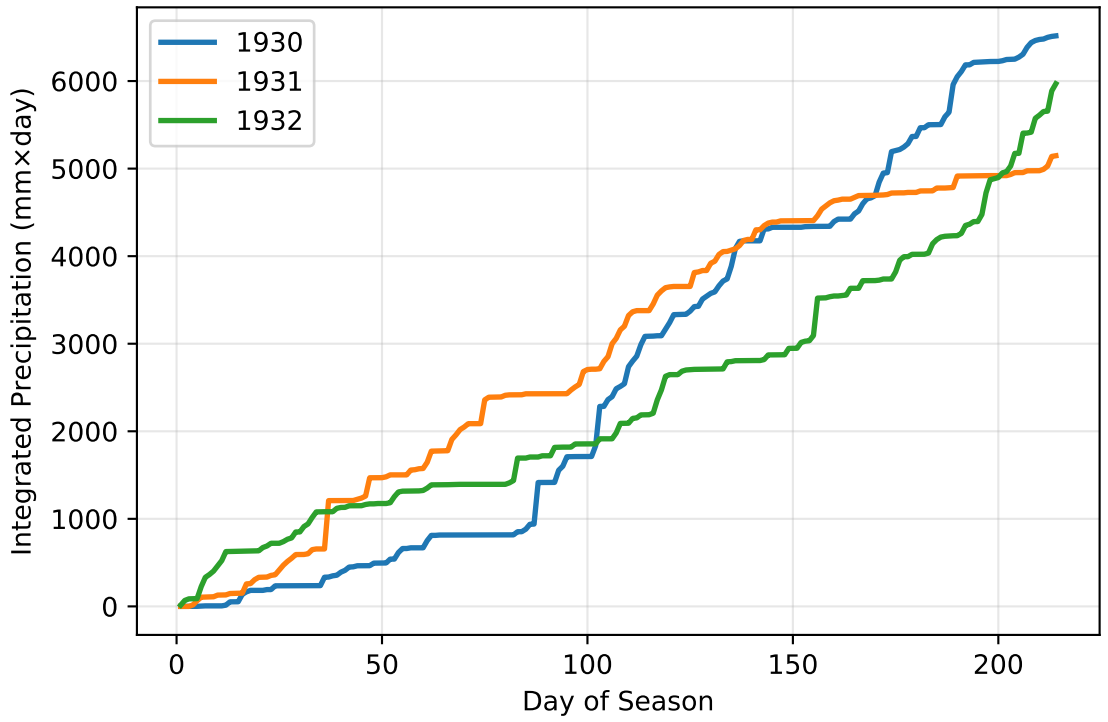

Figure 5.2: Integrated seasonal precipitation patterns. The year 1930 is identified as the wettest, with the highest final integrated value. The year 1931 is the driest, with the lowest integrated value at the end of the season. The smooth curves indicate long-term precipitation behavior.

5.2. Prediction Of Integrated Precipitation Patterns Through SVD

In this section, the results of SVD applied to the integrated precipitation matrix are discussed. The relevant components are selected and the matrix is approximated by the chosen rank. Additionally, the linear prediction is used to forecast the precipitation patterns. The methods used in this section can be found in section 3.3, and section 3.6.

A property of SVD is that the first few singular value values capture a greater part of the variance [16]. This implies that the first few basis vectors are also the most dominant patterns. Figure 5.3 shows the seasonal precipitation patterns. Here, $\vec{v_1}$ decreases over the season. Furthermore, figure 5.4 shows that the coefficient of this seasonal pattern is negative across the years. This means that the pattern shown by the first basis vector is of an increasing precipitation rate. Unfortunately, the coefficients $\vec{u_1}$ and $\vec{u_2}$ do not contribute to additional information corresponding to their respective basis vectors as their signs are mixed. The respective seasonal patterns are similar to each other as they both display a convex behavior, followed by a concave behavior. This highlights the orthogonality property of SVD [26].

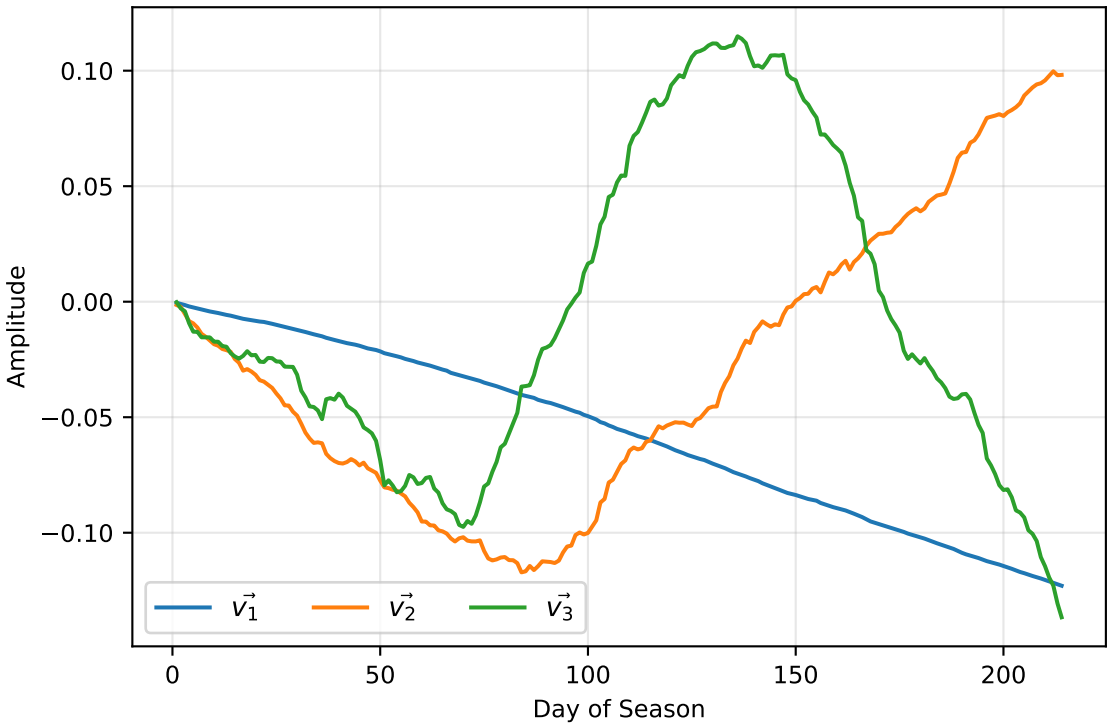

Figure 5.3: The first three basis vectors obtained through singular value decomposition (SVD), representing seasonal precipitation patterns. The first basis vector exhibits a decaying profile over the season, while the second and third display similar structures characterized by convex behavior followed by concavity, indicating shifts in precipitation rates throughout the season.

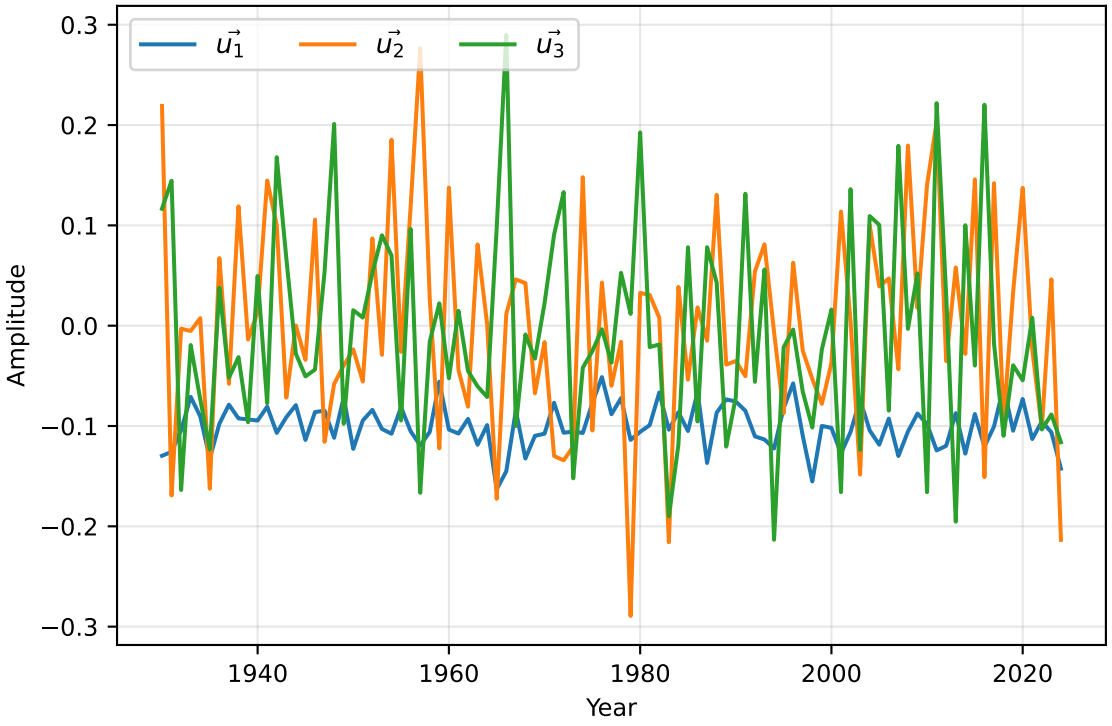

Figure 5.4: Coefficients corresponding to the first three basis vectors. The first coefficient is negative, emphasizing the importance of the pattern represented by the first basis vector. The second and third coefficients exhibit oscillatory behavior.

Figure 5.5 displays the slopes and associated p-values for all basis vectors. The first basis vector exhibits the largest slope, but it is not statistically significant, as its p-value exceeds 0.05 and its coefficient of determination is small as shown in figure 5.6. This apparent discrepancy arises because statistical significance in regression analysis does not necessarily align with the variance captured by singular values. In SVD, the leading singular value explains the majority of the variance in the data, even if the corresponding regression slope is not significant [27]. For this reason, the first basis vector is nevertheless selected for predicting precipitation patterns.

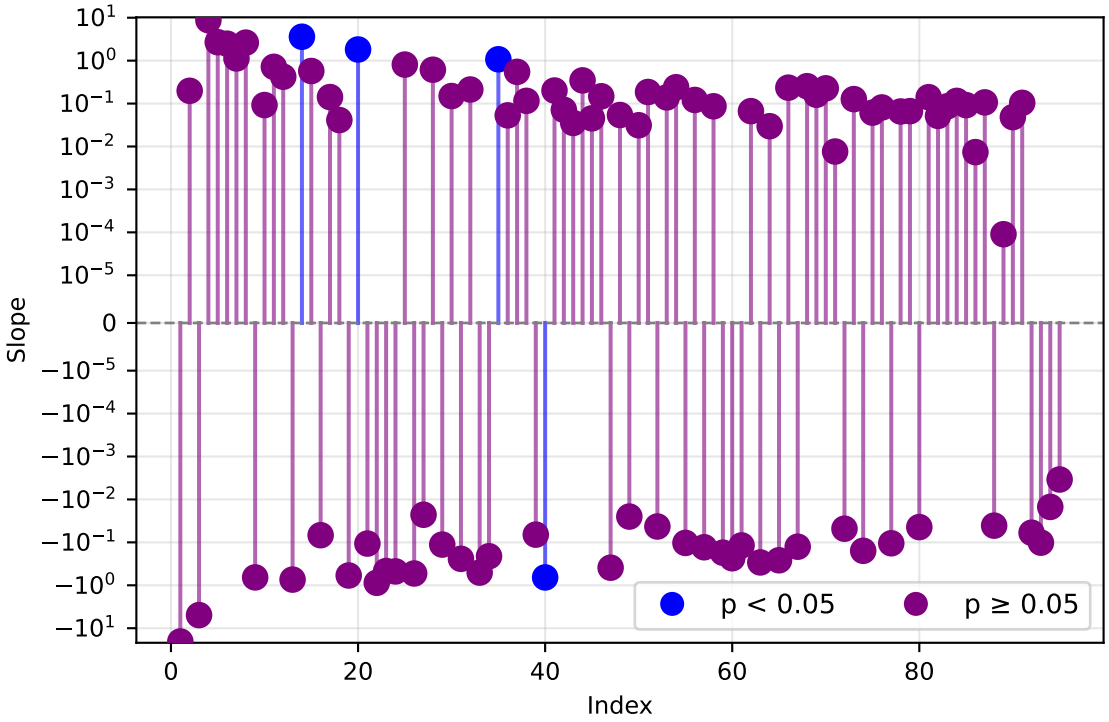

Figure 5.5: Slopes and associated p-values for all components obtained through singular value decomposition (SVD). The first few components are not statistically significant, which contrasts with the expectation from the orthogonality property of the basis vectors.

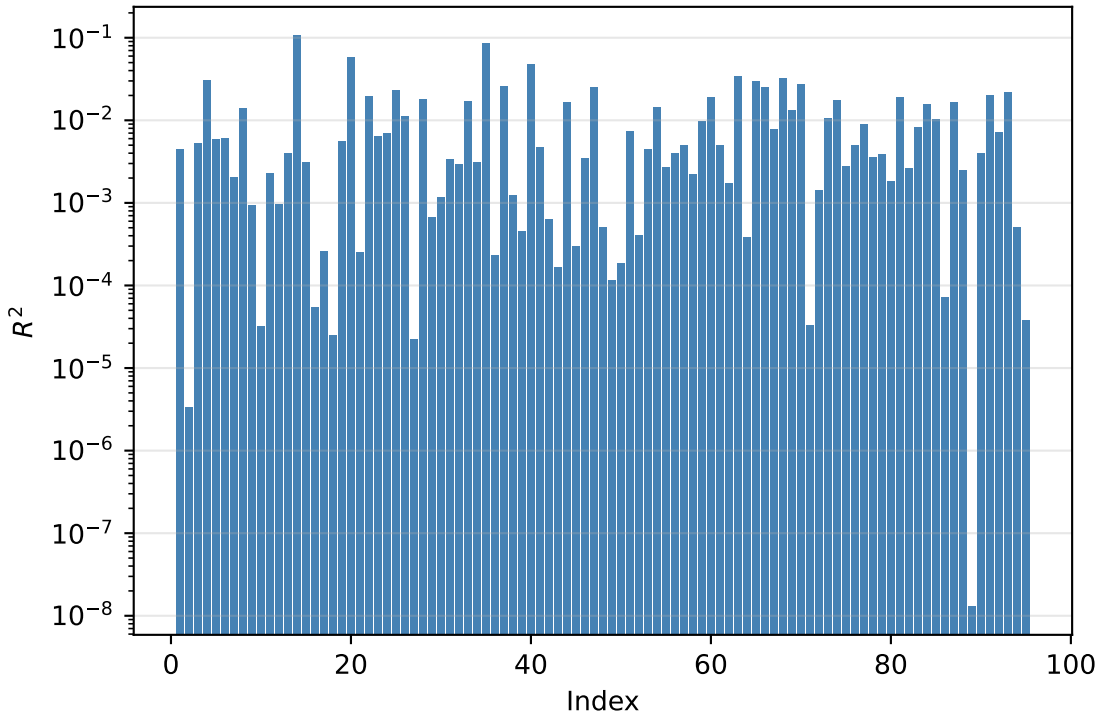

Figure 5.6: Coefficient of determination for all components obtained through singular value decomposition (SVD). All components have values less than or equal to 10^{-1}. This indicates that the components explain only a small fraction of the variance.

Figure 5.7 shows the approximation error for all ranks. This figure illustrates that higher ranks lead to lower approximation errors as more of the data structure is captured. Note that our prediction is restricted to the first rank.

In figure 5.8, the NMSE values of the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$) are presented. The rank-1 approximation initially exhibits a concave NMSE trajectory, reflecting uneven adaptation to precipitation variability. Throughout most of the test years, both methods yield consistently high NMSE values, indicating that precipitation patterns during this period differed substantially from those in the training years. In 2017, precipitation in the Netherlands was relatively evenly distributed across the seasons rather than dominated by extremes. This explains the lower NMSE values observed in that year [28]. After 2017, the two approximations follow a similar trajectory. This results in minor differences in NMSE. In general, the high NMSE values across the test period can be attributed to shifts in precipitation patterns in the Netherlands relative to the training period.

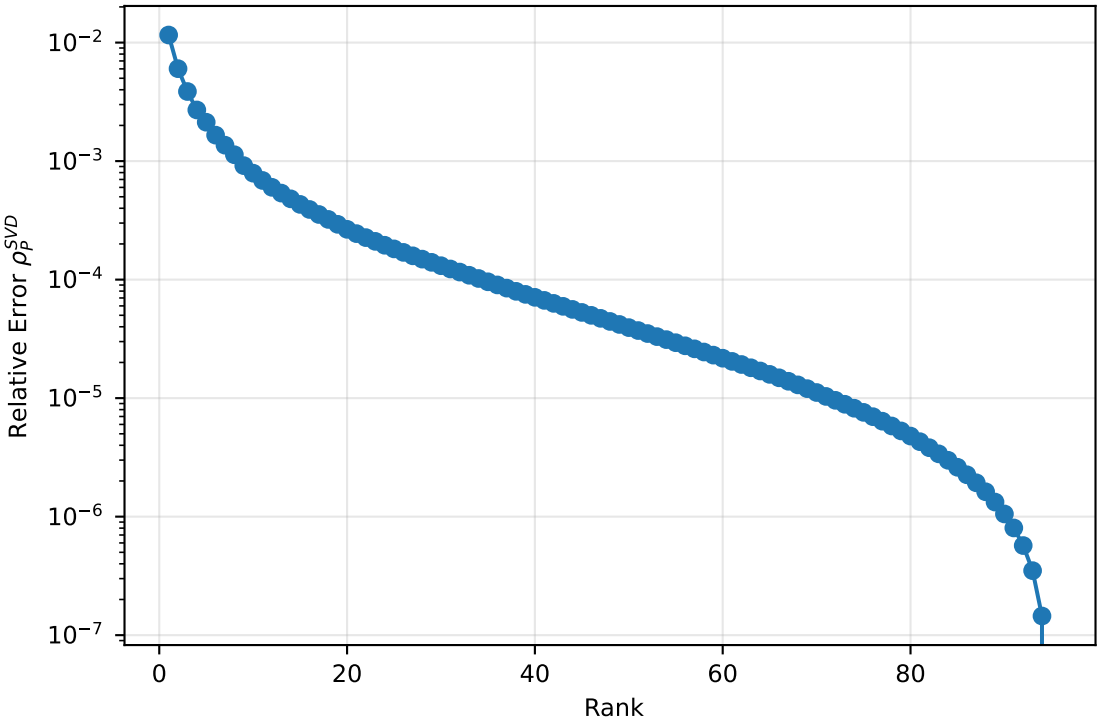

Figure 5.7: SVD approximation error ρ_p^{SVD} as a function of the approximation rank for the integrated precipitation data matrix.

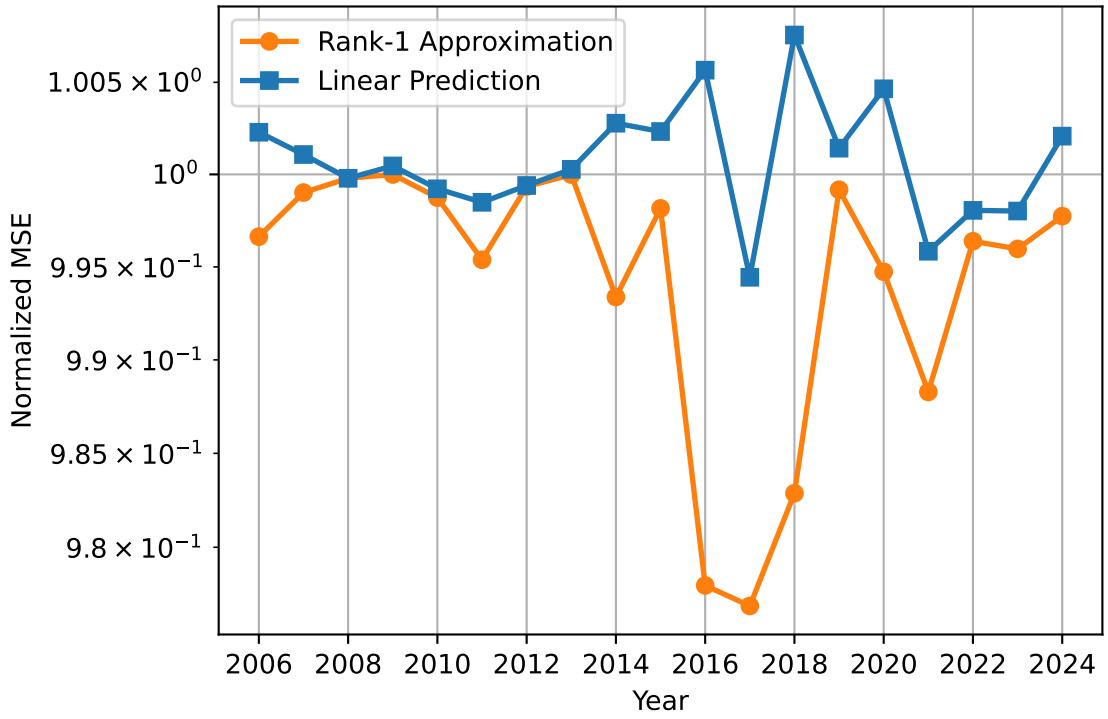

Figure 5.8: Comparison of NMSE values for the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$).

5.3. Forecasting Precipitation Patterns Through NMF

In this section we discuss the results of applying NMF to the integrated precipitation matrix. First, the relative approximation errors are examined across different ranks. Subsequently, statistical measures are analyzed for all NMF components, as these provide insight into their explanatory power. To complement these quantitative results, the most important basis vectors, and their corresponding coefficients are visualized. Finally, the normalized mean squared error (NMSE) values are compared between the rank-1 approximation and the linear prediction model. The methods used for this analysis can be found in section 3.4,and section 3.6,

Figure 5.9 presents the NMF approximation errors ρ_p^{NMF} as a function of the approximation ranks. This figure shows that the NMF algorithm converged for every tested rank, indicating that for each chosen factorization rank, the algorithm reached a stable solution where the reconstruction error no longer decreased significantly. This demonstrates that the integrated precipitation data were well-suited for the applied factorization.

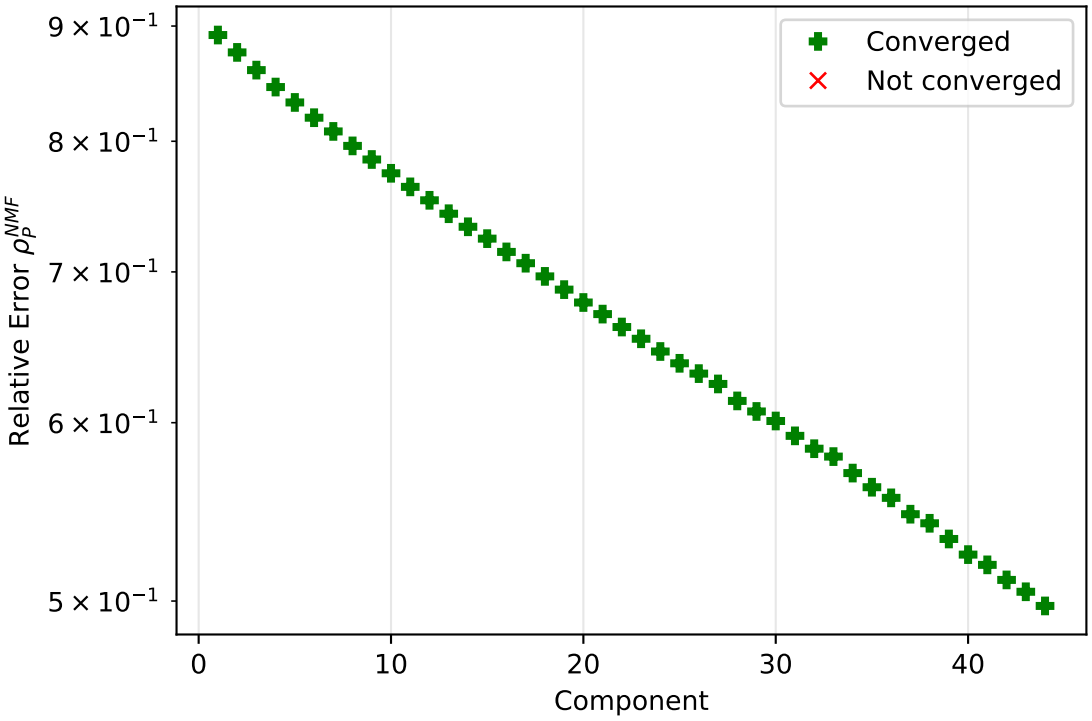

Figure 5.9: NMF Approximation Error ρ_p^{NMF} as a function of the approximation rank for the integrated temperature data matrix. All ranks converged during the NMF algorithm.

Figure 5.10 demonstrates that none of the NMF basis vectors exhibit p-values below the conventional threshold of 0.05. Although the slopes vary in magnitude, particularly for basis vectors with higher indices, they are not statistically significant and therefore cannot be distinguished from random fluctuations. Furthermore, Figure 5.11 presents the coefficients of determination for all basis vectors, with the highest value equal to 0.02, indicating that the NMF decomposition fails to capture long-term precipitation patterns.

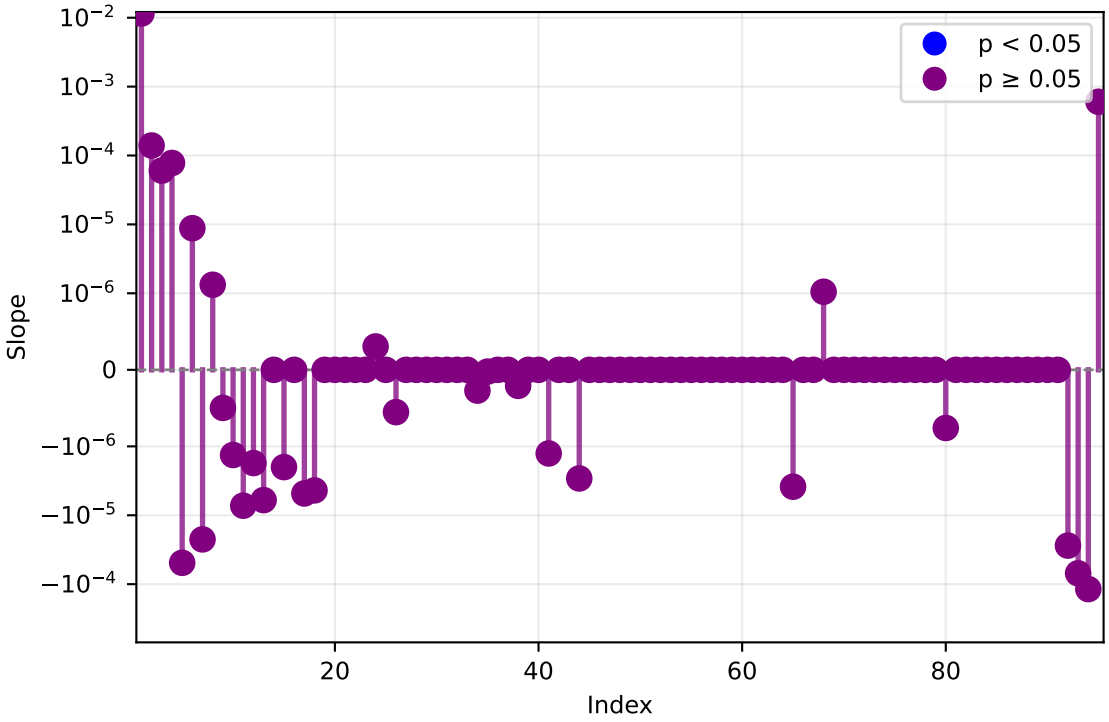

Figure 5.10: Slopes and p-values for all components obtained through NMF. None of these components are statistically significant.

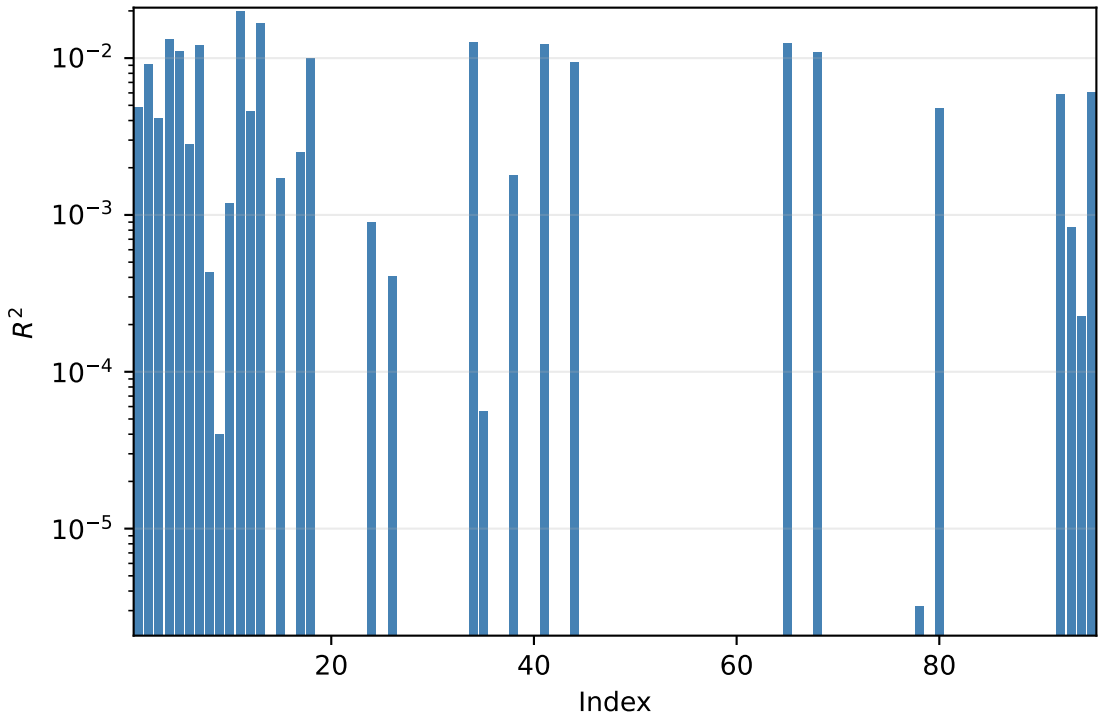

Figure 5.11: Coefficient of determination (R^2) for all the through NMF obtained components. Alle of these coefficients are smaller or equal to 10^{-1}, which signifies a very small explanatory power of each component.

So far, no dominant basis vector has been identified. The behavior of the first three seasonal and

interannual precipitation patterns is shown in figure 5.12 and figure 5.13. For this analysis, the NMF algorithm was rerun with three components. In figure 5.12, the second basis vector resembles the behavior of the first basis vector during the first half of the season, while the third basis vector resembles the behavior of the first basis vector during the latter half of the season. This suggests that the first basis vector captures the overall dominant structure of the precipitation data. The corresponding coefficients remain positive across all years, emphasizing the consistent importance of the precipitation patterns. Consequently, further predictions of precipitation patterns will be based on the first basis vector.

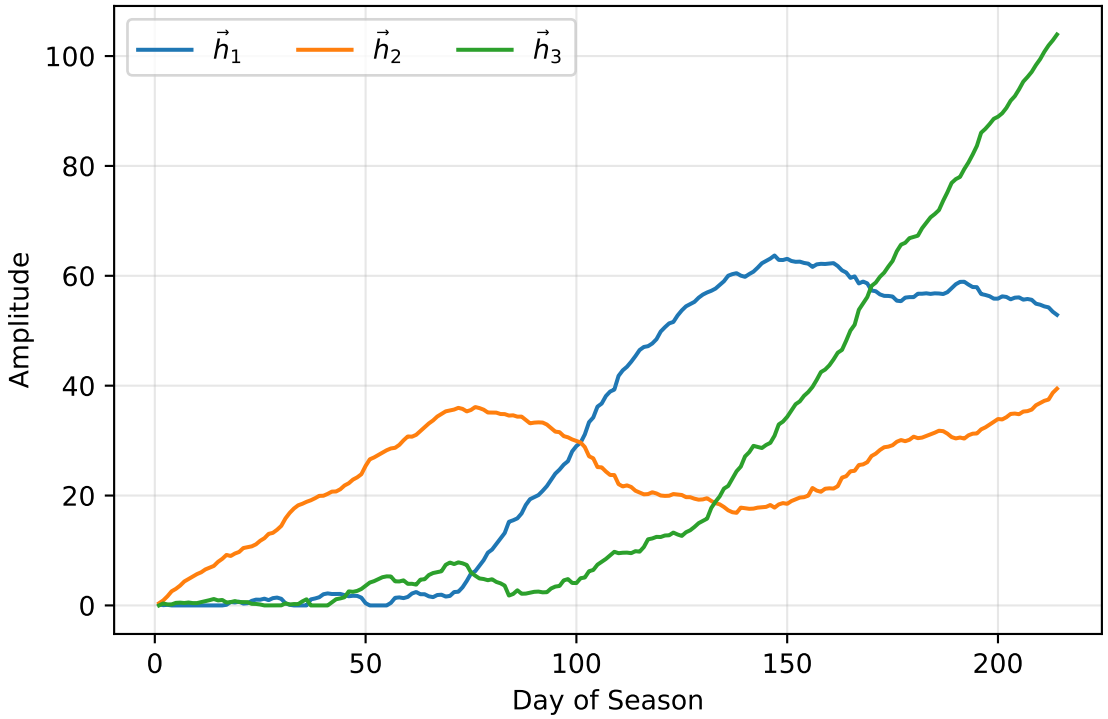

Figure 5.12: The three NMF basis vectors obtained by using three components during the initialization of the algorithm. These are the seasonal precipitation patterns.

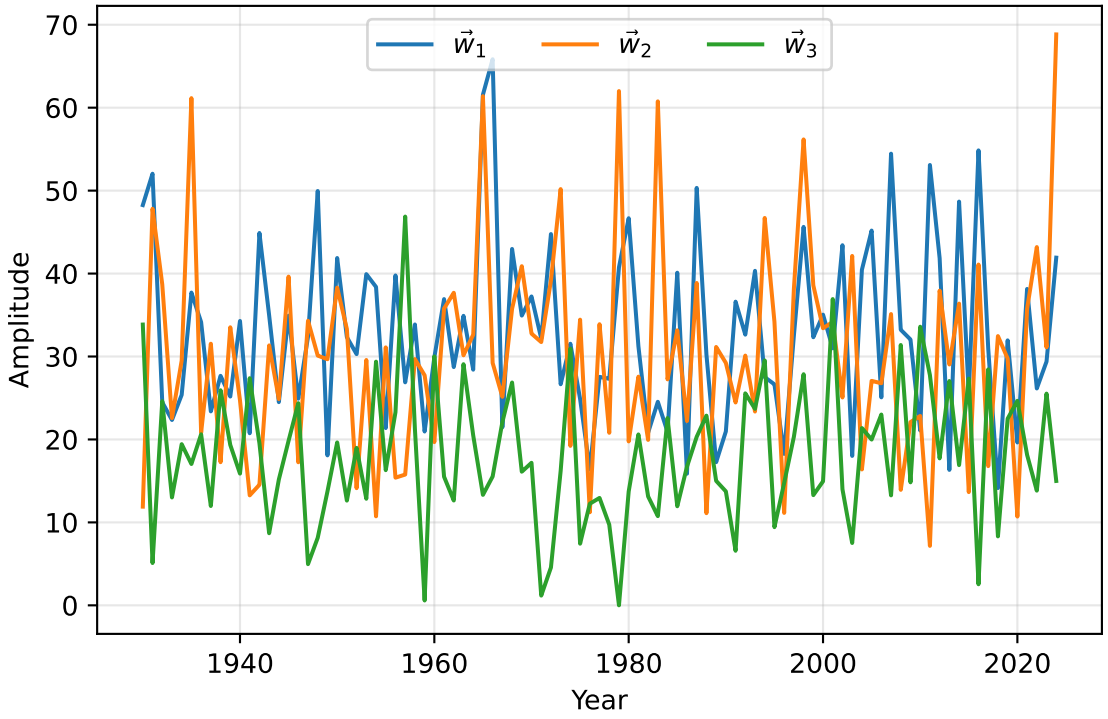

Figure 5.13: The coefficients corresponding to the basis vectors obtained through NMF. These are the interannual precipitation patterns. All these coefficients are positive, contributing positively to the relevance of the corresponding seasonal precipitation pattern.

Building on this reasoning, predictions are carried out using the first basis vector. The performance of these predictions is shown in figure 5.14, which compares the normalized mean squared error (NMSE) values of the rank-1 approximation and the linear prediction model. The two approaches produce nearly identical results, with only minor differences in error magnitude. Notably, the error peaks are smaller than those observed in figure 4.8, indicating that both methods capture the structure of the NMF-derived basis vector effectively. The minor differences between the NMSE values confirms that the first basis vector provides a stable and reliable representation of the precipitation patterns.

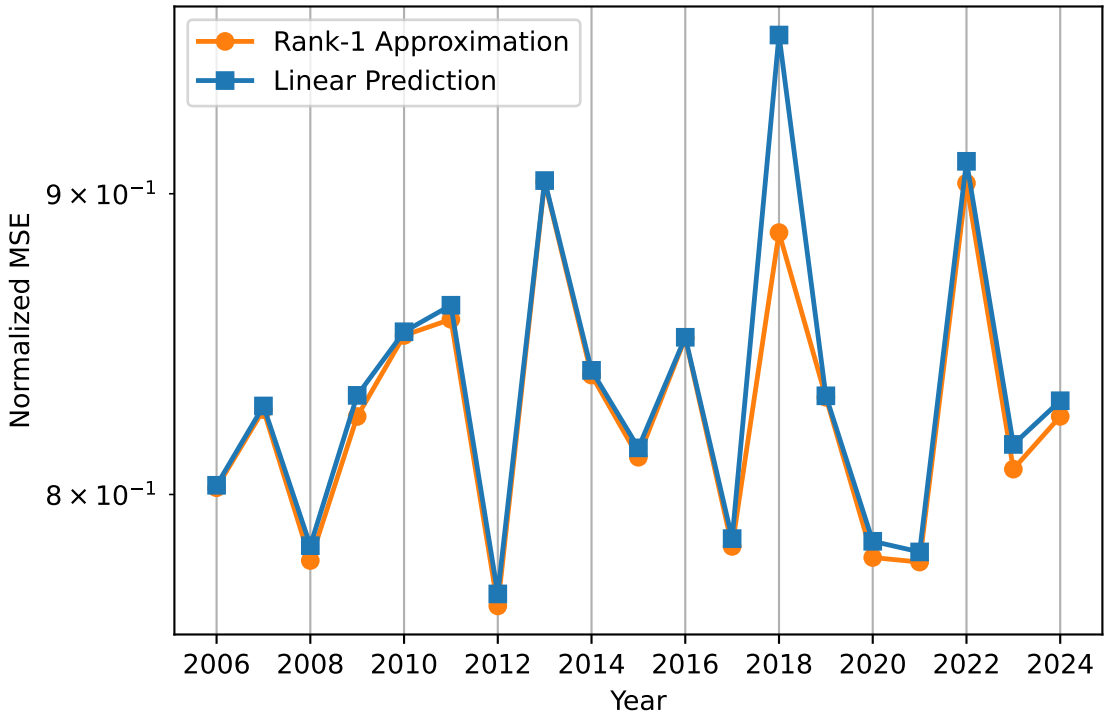

Figure 5.14: Comparison of NMSE values for the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$). The linear prediction model captured the structure of the precipitation patterns well based on the overlapping NMSE values.

6

The Hydrothermal Interaction

This chapter focuses on predicting hydrothermal interaction patterns, which reflect the coupled dynamics of temperature and precipitation. Changes in these interactions directly affect soil moisture balance, crop growth, and water resource availability. In the Netherlands, agriculture and flood management are highly sensitive to both heat and rainfall, emphasizing the importance of forecasting hydrothermal interaction patterns. This chapter examines long-term hydrothermal interaction patterns by analyzing the integrated precipitation matrix, which emphasizes cumulative seasonal behavior rather than short-term variability in section 6.1. Singular value decomposition (SVD) and non-negative matrix factorization (NMF) are applied to identify dominant temporal components and to evaluate their predictive value in section 6.2,and section 6.3. This provides the basis for assessing which representation offers the most robust characterization of hydrothermal interaction dynamics in the Netherlands.

6.1. Visualization Of The Hydrothermal Data matrices

Figure 6.1 presents the first three rows of the hydrothermal matrix. These patterns closely resemble those of the precipitation matrix, with the distinction that the peaks are reduced in magnitude. This observation indicates that precipitation is the primary driver of the hydrothermal interaction term.

Furthermore, figure 6.2 illustrates the cumulative hydrothermal patterns for the years 1981–1983. In all cases, the curves begin with relatively low values at the start of the season, increase steadily toward mid-season, and subsequently decline toward the end of the season. This reflects a hydrothermal cycle. Despite this similarity in shape, the amplitudes differ across years. The year 1981 exhibits the highest peak, indicating stronger hydrothermal conditions. Whereas 1983 remains consistently lower, reflecting weaker interaction.

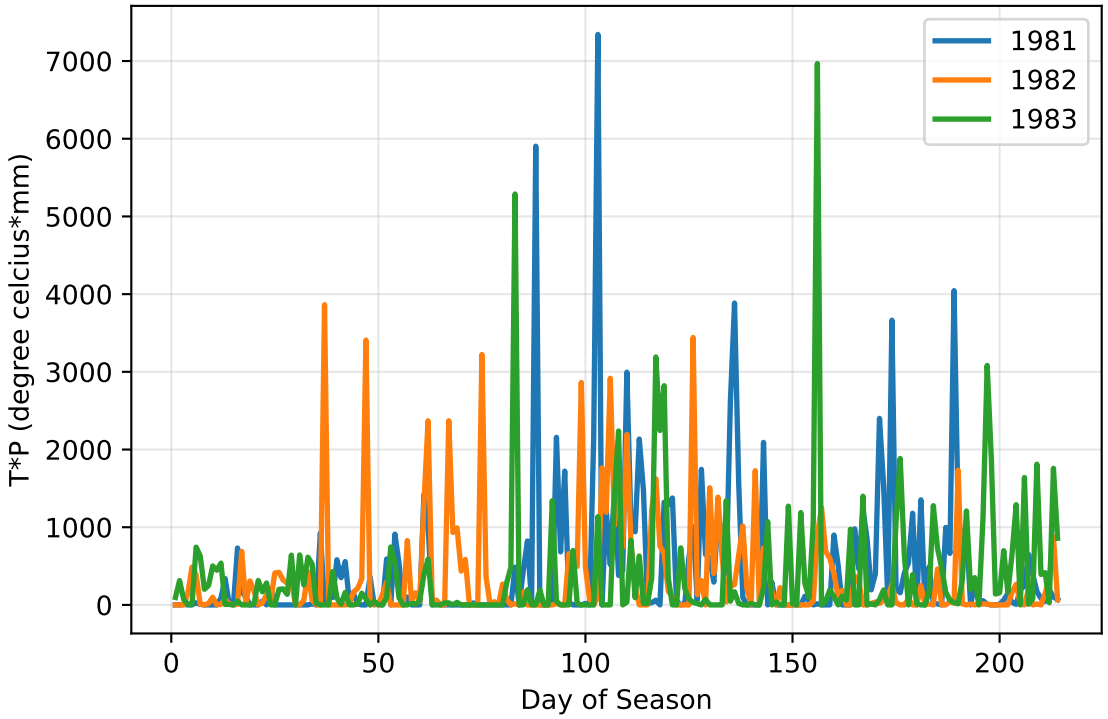

Figure 6.1: First three rows of the hydrothermal matrix. This is similar to figure 5.1. However, the peaks are smaller.

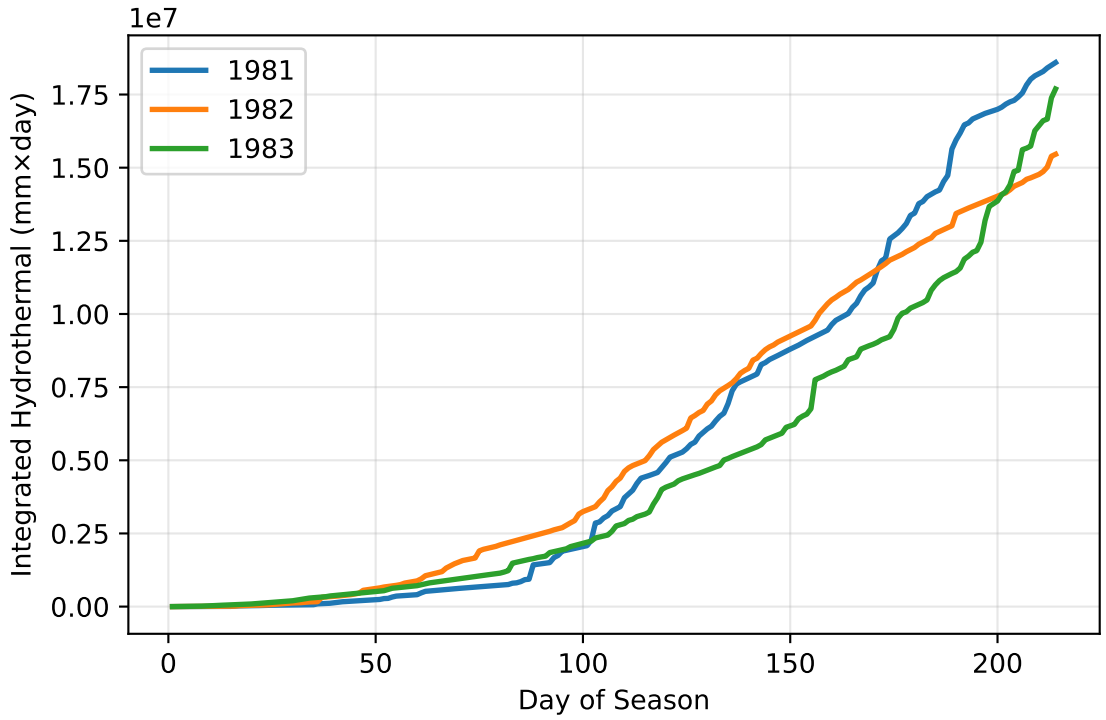

Figure 6.2: Seasonal trajectories of the integrated hydrothermal matrix for the first three years (1981–1983).

6.2. Prediction Of The Integrated Hydrothermal Patterns Through SVD

This section presents the results of applying singular value decomposition (SVD) to the integrated hydrothermal interaction data matrix. The analysis begins with an evaluation of the relative approximation error. The extracted basis vectors and their associated coefficients are then examined to identify the dominant structures underlying the data. The statistical relevance of these components is assessed through the analysis of ,p-values, and coefficients of determination (R^2). Finally, predictive performance is evaluated by comparing normalized mean squared error (NMSE) values obtained from both rank-1 approximation ,and linear prediction. These elements establish a systematic framework for assessing the extent to which SVD captures and predicts the integrated hydrothermal dynamics. The methods used for this analysis can be found in section 3.3, and section 3.6.

Figure 6.3 shows the relative approximation error as a function of the rank. As expected, the error decreases when more ranks are included in the reconstruction of the data matrix. However, the figure makes clear that the majority of this reduction occurs within the first three ranks. Beyond this point, additional ranks yield only marginal improvements. This observation highlights that the essential structure of the integrated hydrothermal interaction matrix is captured by the first three basis vectors.

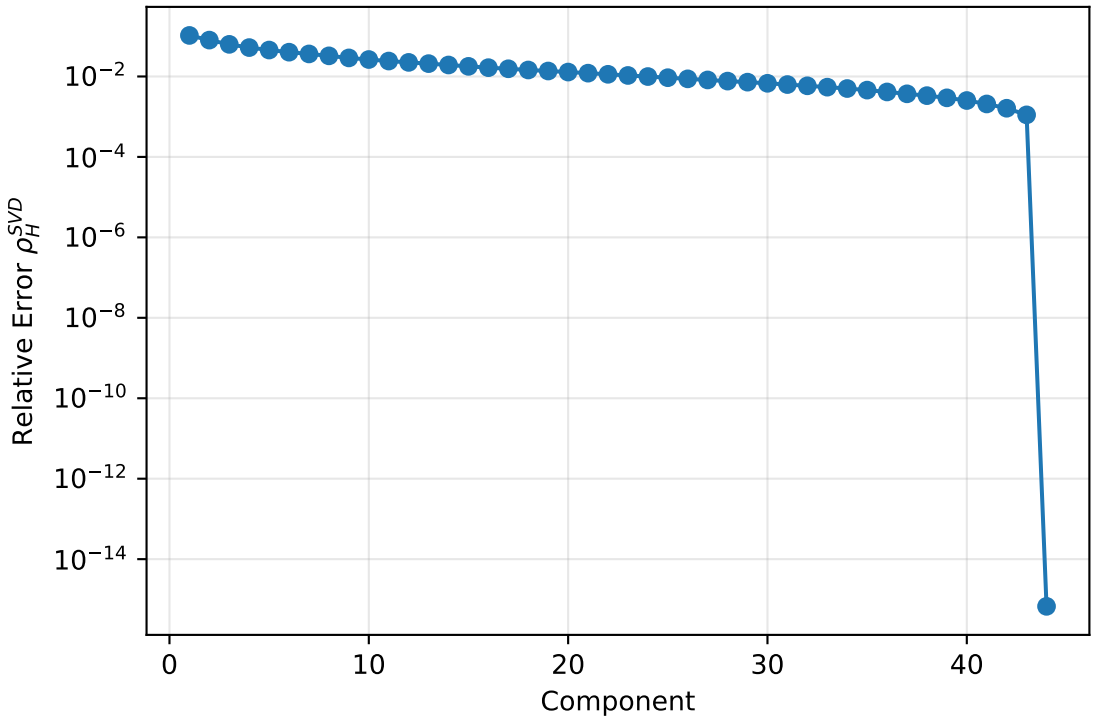

Figure 6.3: Relative approximation error ρ_H^{SVD} as a function of approximation rank. The error decreases substantially with the first three ranks.

Figure 6.4 shows the first three basis vectors, also referred to as the first three hydrothermal interaction patterns across the season. The first basis vector captures the dominant seasonal trend: a gradual increase in hydrothermal flux toward peak summer, reflecting the combined influence of rising temperatures and sustained precipitation. The second and third basis vectors exhibit oscillatory behavior within the season. These components represent residual variability rather than physically interpretable signals.

Figure 6.5 displays the three coefficients corresponding to the basis vectors. The first coefficient remains consistently negative across years, indicating the constant relevance of the pattern captured by the first basis vector. By contrast, the second and third coefficients exhibit oscillatory behavior. The

latter two coefficients contribute little to no interpretive value beyond their associated basis vectors.

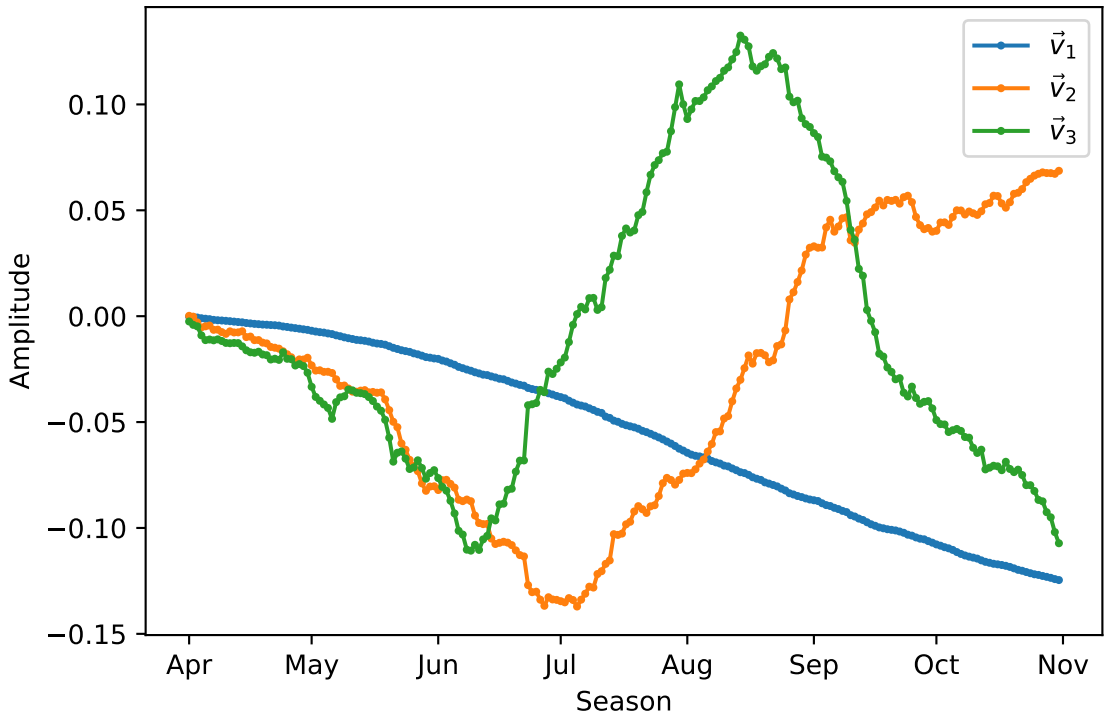

Figure 6.4: The first three basis vectors generated through SVD. These describe the hydrothermal interaction patterns within the season.

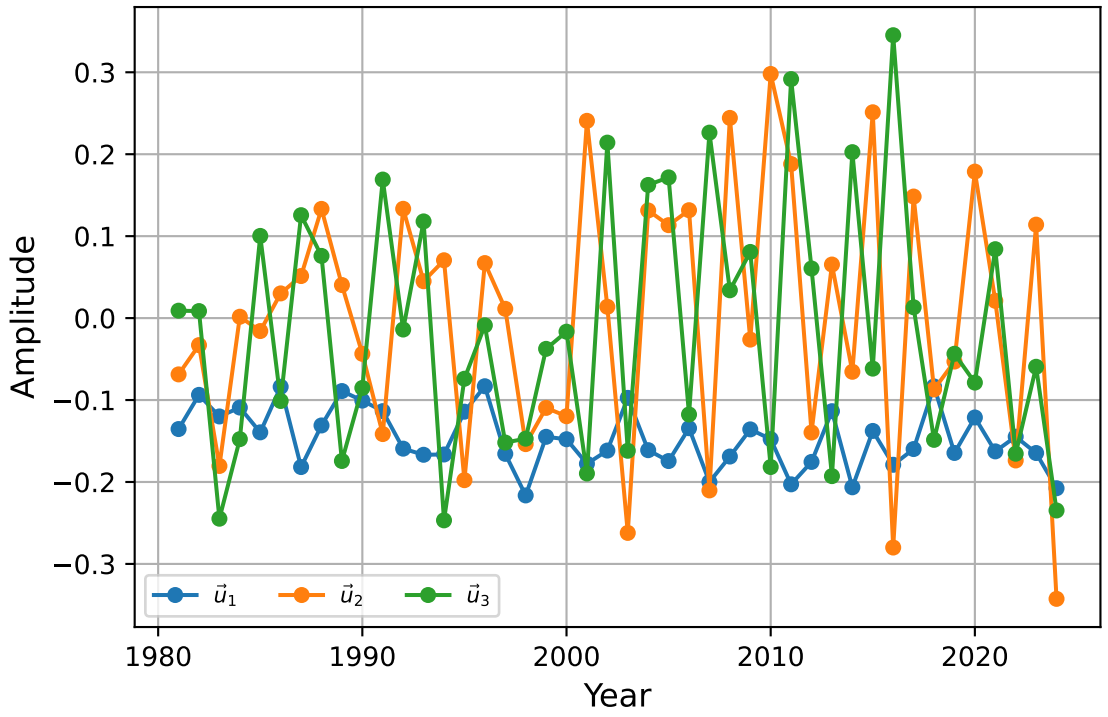

Figure 6.5: The first three coefficients, which correspond to the basis vectors. These are the interannual hydrothermal interaction patterns.

Figure 6.6 displays the estimated slopes for all components obtained through SVD. The components that exhibit the largest slopes are not statistically significant. However, the components with the smallest slopes are statistically significant, highlighting the importance of statistical testing when identifying reliable trends. As shown in figure 6.7, the first component attains the second-highest coefficient of determination (R^2). Due to the orthogonality property of SVD, the first component captures the greatest explanatory variance, making it the most suitable candidate for predictive analysis.

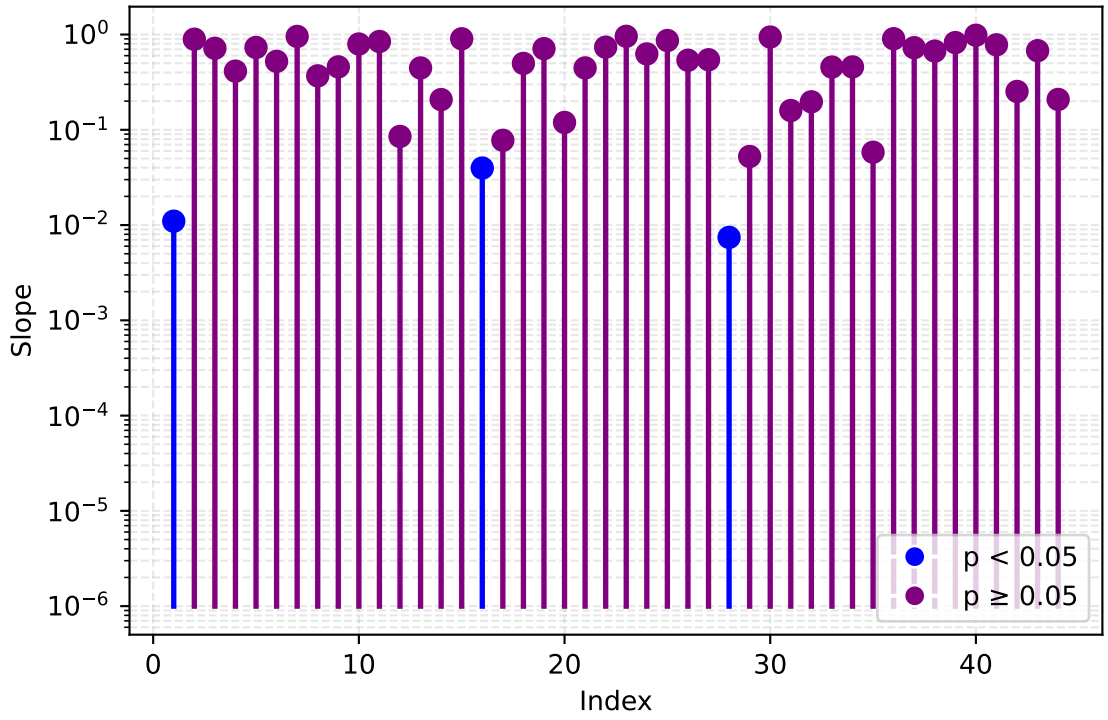

Figure 6.6: Slopes and associated p-values for all components obtained through SVD. Statistically significant components exhibit slopes that are relatively smaller in magnitude.

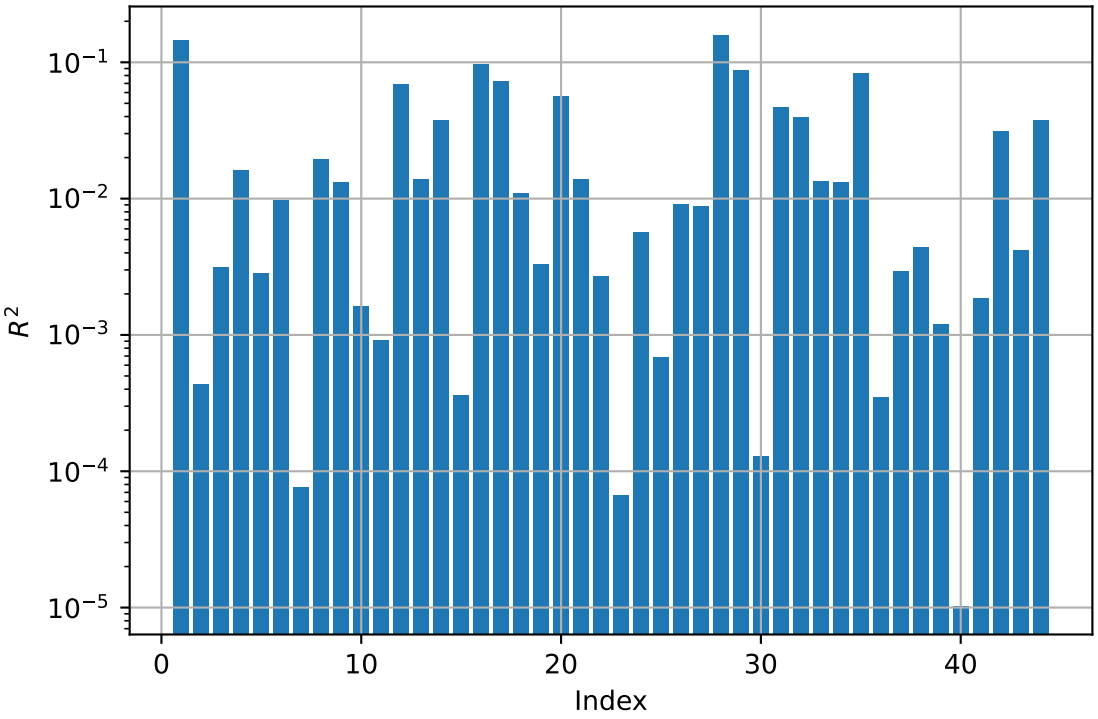

Figure 6.7: Coefficient of determination R^2 for each of the components obtained through SVD. The statistically significant components from figure 6.6 exhibit the highest R^2-values.

In figure 6.8, the NMSE values of the rank-1 approximation, $NMSE_{\text{approx}}(t_{\text{tst}})$, and the linear prediction, $NMSE_{\text{pred}}(t_{\text{tst}})$, are shown. Overall, the linear prediction consistently exhibits higher NMSE values than the rank-1 approximation. Furthermore, three notable irregularities can be identified.

In 2018, the Netherlands experienced an exceptional drought combined with unusually high temperatures in the summer. This sudden change in temporal and precipitation patterns affected the hydrothermal interaction term [2]. The rank-1 approximation achieved a very small NMSE compared to the linear prediction. This indicates that the dominant spatio-temporal component captured by SVD could adapt to these conditions, unlike the linear prediction.

A similar situation arose in 2020, one of the warmest years on record in the Netherlands. Characterized by a long August heatwave, very dry spring conditions, and simultaneously extreme precipitation in February [29]. These contrasting events complicated the hydrothermal interaction term. While the rank-1 approximation could still capture the dominant structure, the linear prediction failed to represent the irregular combination of extremes.

The third peak occurred in 2022. Both models obtained high NMSE values. That year was exceptionally warm, and extremely dry, with severe precipitation deficits [4]. These conditions disrupted the hydrothermal interaction term to such an extent that even the rank-1 approximation could not maintain low error. This explains why in 2022 both methods performed poorly, unlike in earlier peaks where the rank-1 approximation still retained smaller NMSE values.

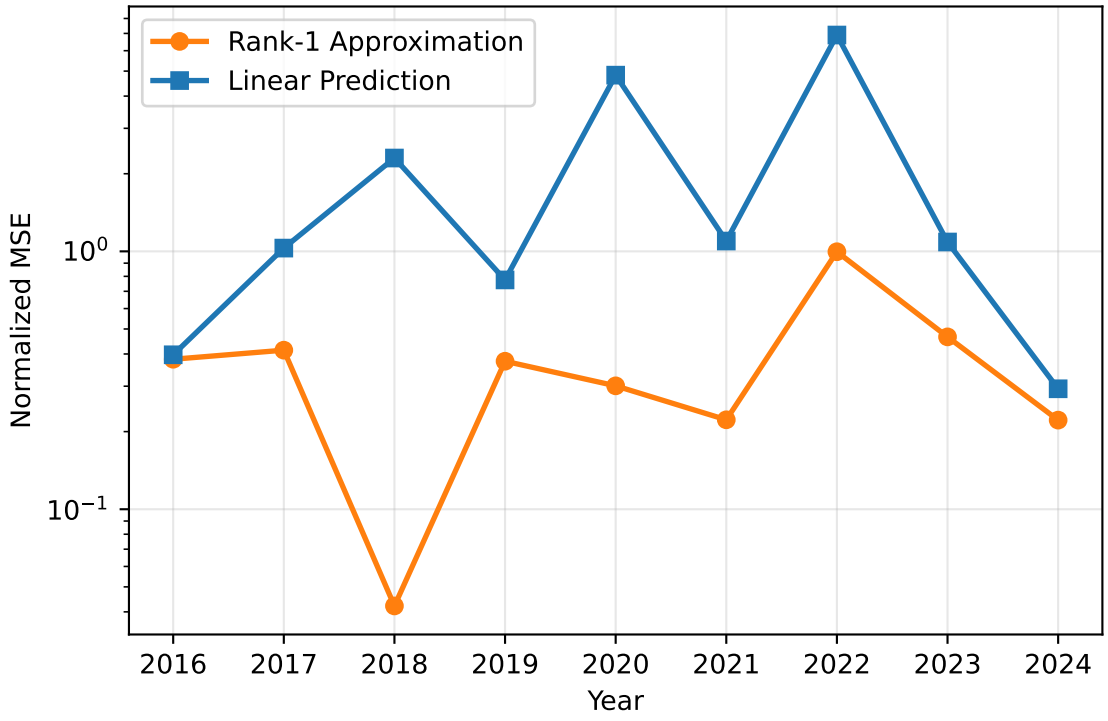

Figure 6.8: Comparison of NMSE values for the rank-1 approximation ($NMSE_{approx}(t_{tst})$) ,and the linear prediction ($NMSE_{pred}(t_{tst})$). The linear prediction model had overall higher NMSE values. Three irregularities occurred in 2018,2020, and 2022, which had consequences for the NMSE values of both models.

6.3. Prediction of Integrated Hydrothermal Patterns using NMF

The analysis used in this section is similar to the previous section, differing in the fact that it focuses on NMF instead of SVD.

Figure 6.9 plots the relative approximation error as a function of the approximation rank. As expected, increasing the number of components reduces the error. However, convergence is only achieved for the first nine ranks, while higher-order components fail to stabilize, as indicated by the red ticks. This highlights a key limitation of NMF: unlike SVD, it does not guarantee orthogonality or variance-ordering of components. This makes it difficult to identify which basis vectors are most critical for reconstructing the hydrothermal interaction matrix. To overcome this limitation, we further examine the slopes, associated p-values, and coefficients of determination of all components.

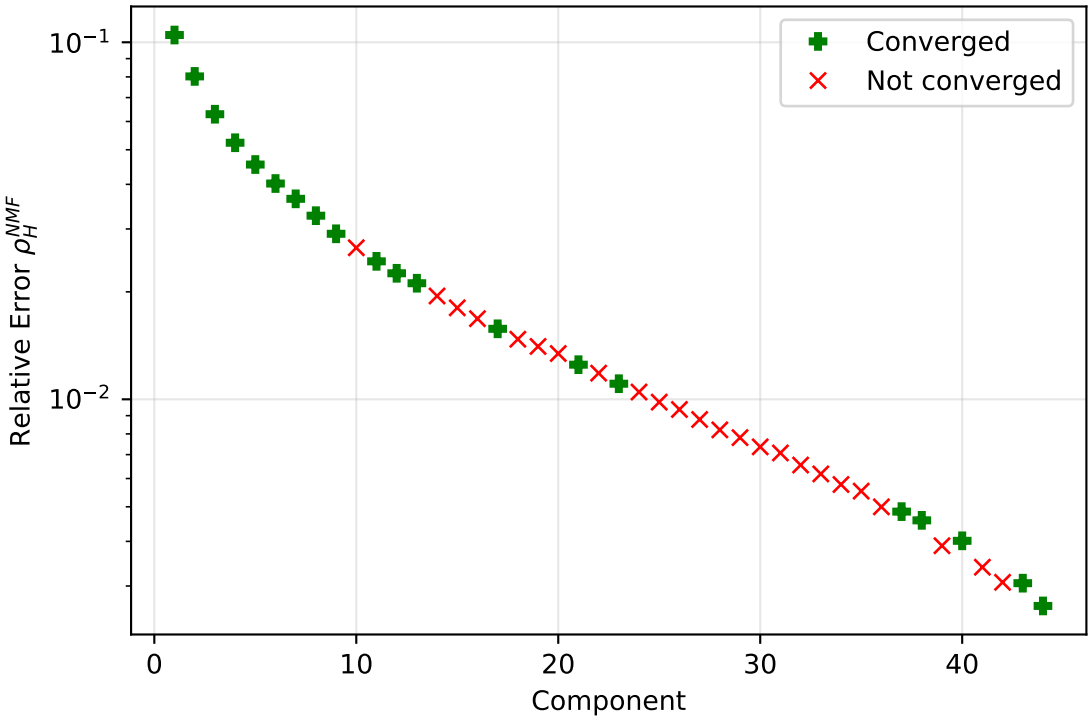

Figure 6.9: Relative Approximation Error ρ_H^{NMF} as a function of the approximation ranks. The first nine ranks were able to converge during the NMF algorithm. Overall the components with higher indices were not able to converge during the NMF algorithm.

Figure 6.10 displays the slopes and associated p-values for all NMF components. The first component exhibits the largest slope and is statistically significant. While the three higher-indexed components $14, 25, 42$ also achieve significance despite their smaller magnitudes. As shown in figure 6.11, all components obtained coefficients of determination below 10^{-2}, indicating that no component provides strong explanatory power. This contrast highlights another key limitation of NMF: statistical significance in slope estimates does not necessarily translate into meaningful variance explanation. Consequently, the first component is selected, as it combines statistical significance with a relatively high explanatory strength compared to the other components.

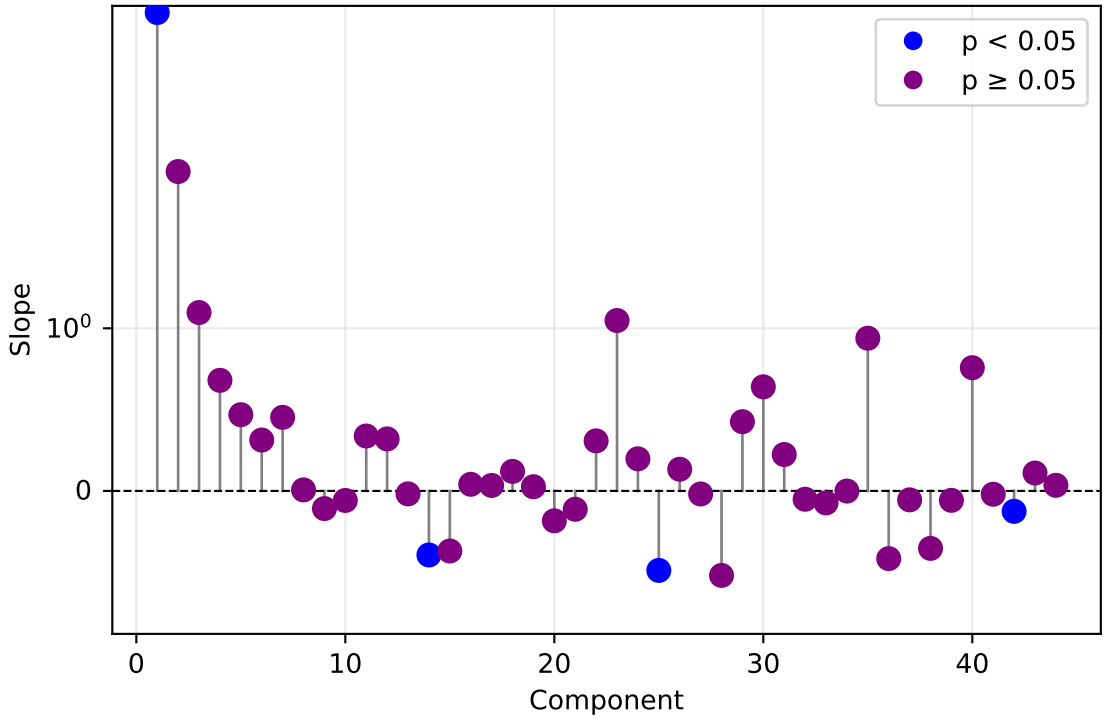

Figure 6.10: Slopes and p-values for all components obtained through NMF. The first component exhibits the largest slope and p-value less than 0.05. Higher-indexed components have smaller slopes and are not statistically significant, except for components $14, 25, 42$.

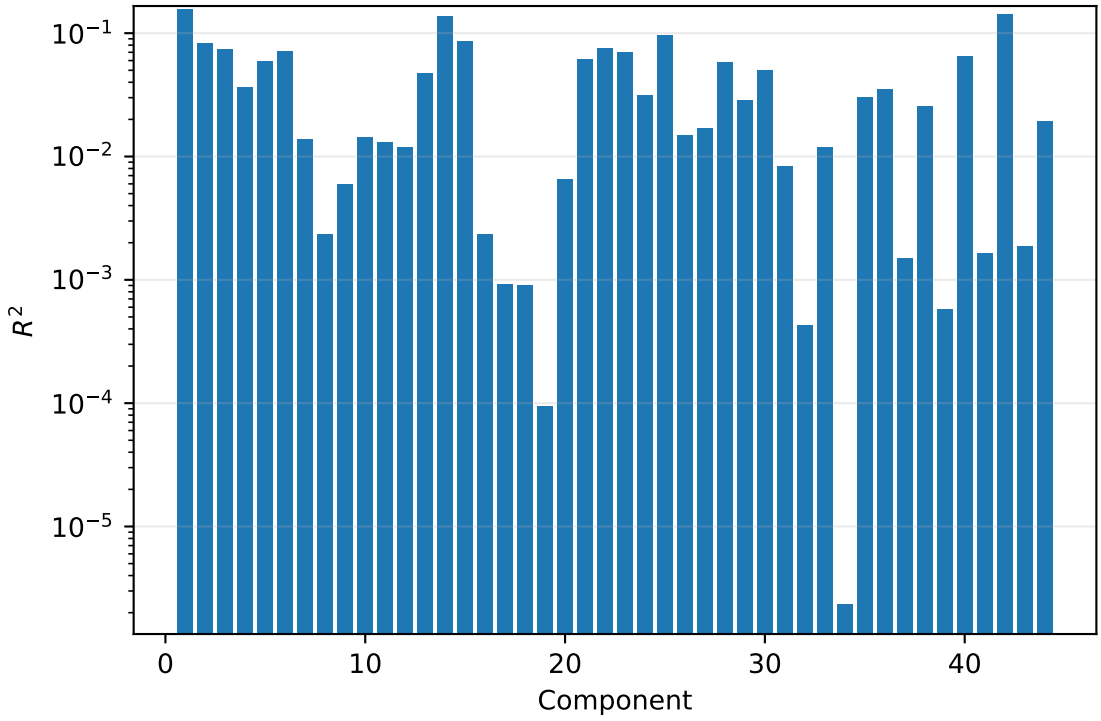

Figure 6.11: Coefficient of Determination (R^2) for all components obtained throough NMF. All components exhibit $R^2 < 10^{-2}$. The first component has the highest R^2-value. This translates into a relatively high explanatory power.

Figure 6.12 shows the first three basis vectors obtained from re-running the NMF algorithm with three components. While the second and third vectors display oscillatory behavior, the first basis vector captures a complete seasonal pattern. This component exhibits a mid-season rise, consistent with intensified precipitation. The corresponding coefficients in figure 6.13 remain positive across all years.

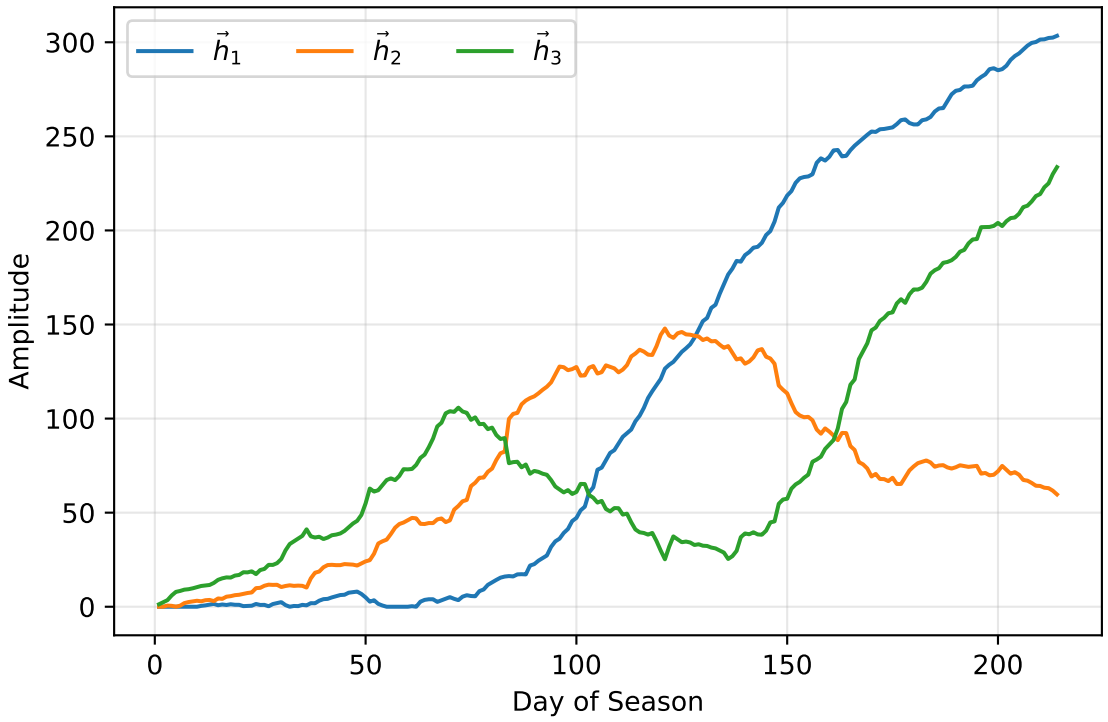

Figure 6.12: The first three basis vectors obtained from re-running the NMF algorithm with three components. These basis vectors capture the dominant seasonal hydrothermal interaction patterns.

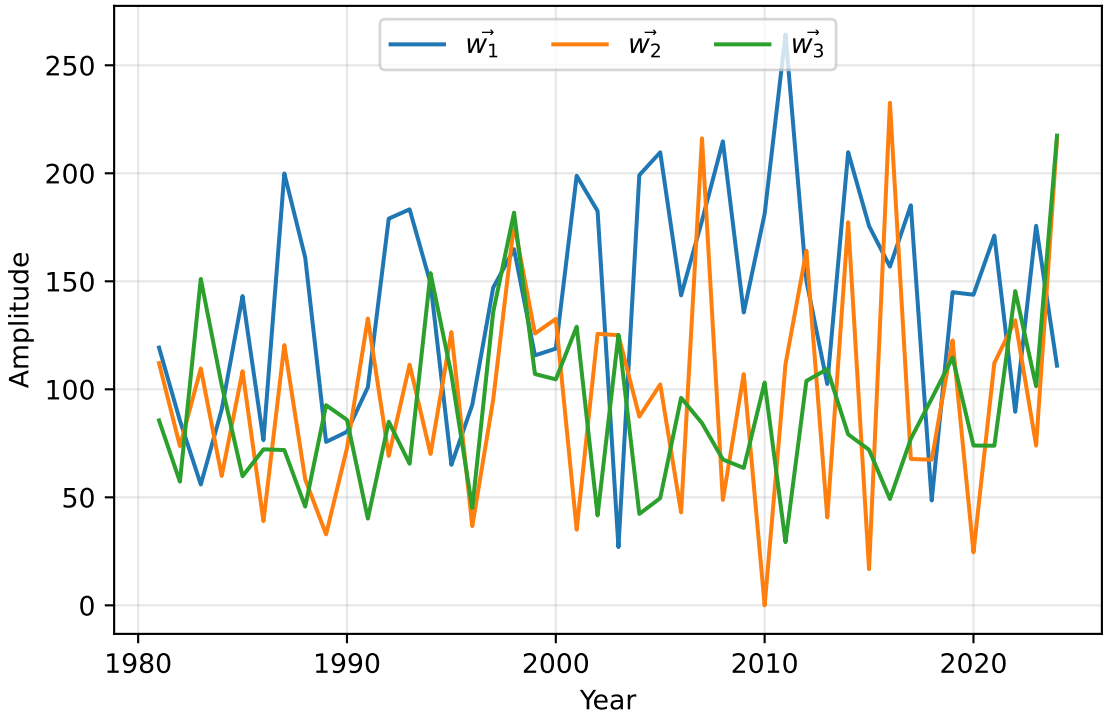

Figure 6.13: Coefficients associated with the first three basis vectors obtained through NMF. These coefficients illustrate how the hydrothermal patterns evolve across years.

In figure 6.14, the NMSE values of the rank-1 approximation and the linear prediction model are presented. The rank-1 approximation consistently achieves lower NMSE values. The linear prediction curve lies above it throughout the test period. Moreover, the linear prediction exhibits successive peaks and troughs. This indicates instability in capturing the underlying hydrothermal interaction patterns. This highlights the superior performance and robustness of the rank-1 approximation for representing the hydrothermal interaction data.

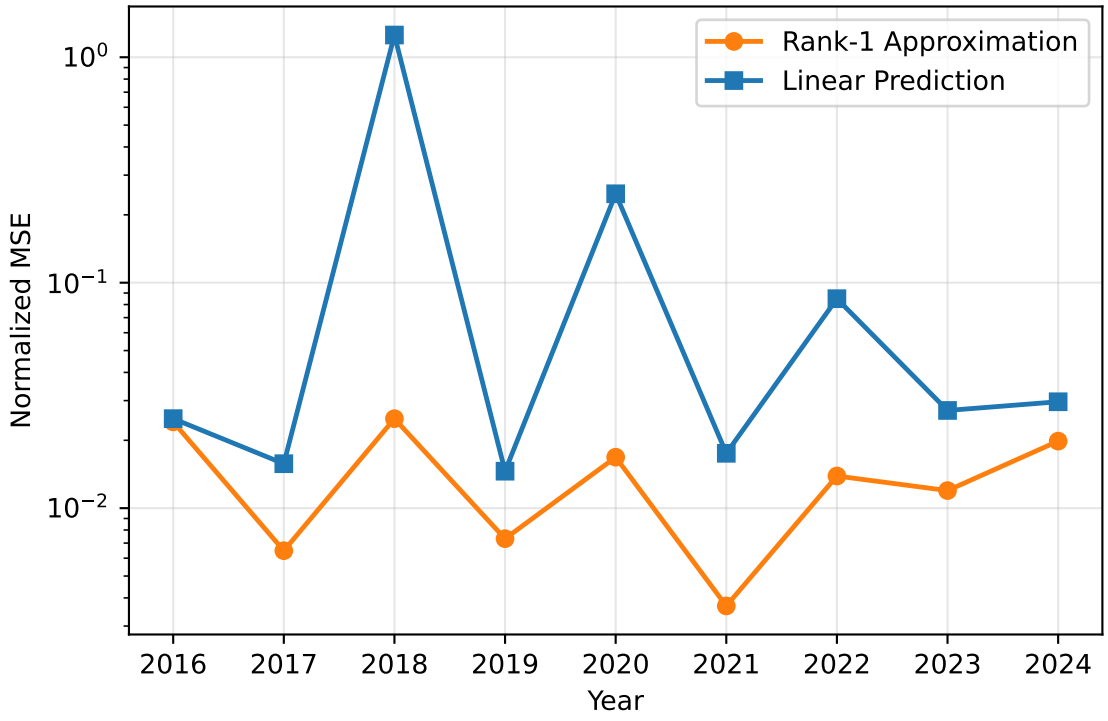

Figure 6.14: Normalized mean squared error values for the rank-1 approximation $NMSE_{approx}(t_{tst})$, and the linear prediction model $NMSE_{pred}(t_{tst})$. The errors of the linear prediction consistently lie above those of the rank-1 approximation.

7

Conclusion

This study addressed the following research question: **To what extent do SVD-based and NMF-based representations, in combination with linear prediction models and rank approximations, capture seasonal and interannual variability in temperature, precipitation, and their hydrothermal interactions in the Netherlands?**. Here, the term hydrothermal interaction term reflects the coupled dynamics of temperature and precipitation.

This study used two decomposition techniques: singular value decomposition, and non-negative matrix factorization. These techniques were applied to the temperature, precipitation, their correspponding integrated matrices, and hydrothermal interaction data matrices. The approximation errors were plotted against the ranks to gain insight into how many components were needed to capture the seasonal and interannual variations. For the environmental variables studied, the first component captured the seasonal and interannual variations. The linear prediction and rank-1 approximation were used to forecast patterns of these environmental variables, which were assessed by means of the Normalized Mean Squared Error.

The research question regarding which predictive framework can adequately capture seasonal variations in temperature, precipitation, and their hydrothermal interactions in the Netherlands remains only partially resolved. Nevertheless, the seasonal patterns obtained through Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF) provided valuable insights.

For temperature, SVD extracted a broad seasonal trend, but its orthogonality constraint produced patterns that were not realistic. In contrast, NMF yielded additive and consistently positive components, with the first component representing the dominant seasonal variation and higher-indexed components accounting for residual fluctuations. For precipitation, SVD concentrated most of the variance in the first component, suggesting an overall seasonal increase, while later components offered limited interpretability. NMF provided a clearer structure: the first component captured the dominant seasonal signal, and the remaining components reflected secondary variation. The positive coefficients across years indicates that NMF offers a more reliable representation of precipitation dynamics. For hydrothermal interactions, SVD identified a combined seasonal trend reflecting rising temperatures and sustained precipitation, but additional components mainly captured oscillatory noise. On the other hand, NMF produced a seasonal pattern that remained consistently relevant across years, emphasizing its suitability for complex variables such as the hydrothermal interaction term.

The predictive framework that most effectively captured interannual temporal variations was the rank-1 approximation using the component obtained through NMF. Across the testing dataset, which comprised 20% of the full data, rank-1 approximations consistently outperformed linear predictions. In the case of temperature, rank-1 approximation derived with the component obtained through NMF exhibited smaller error peaks and more stable NMSE values across the years. This is an example of its superior ability to preserve underlying seasonal components and adapt to shifts in dynamics. For precipitation, both rank-1 approximation (with the component obtained through SVD) and linear prediction yielded consistently high NMSE values. There was a temporary improvement in 2017. On the other

hand, the rank-1 approximation (with the component obtained through NMF) produced nearly identical NMSE values to the linear prediction. Hydrothermal interactions highlighted the contrast between models, with the rank-1 approximation (with the component obtained through SVD) outperforming linear prediction during extremes in climate such as in the years 2018 and 2020. It adapted to dominant spatio-temporal component, yet failed under irregular conditions in 2022. By comparison, the rank-1 approximation (with the component obtained through NMF) consistently achieved lower NMSE values, emphasizing its robustness as a reliable predictive framework for hydrothermal dynamics.

Future research could benefit by incorporating a wider testing dataset size, and a range of predictive approaches. Thereby enabling more robust comparisons across methods. Furthermore, these predictive frameworks can also be applied to systematically evaluate which model best captures seasonal variations in temperature, precipitation, and hydrothermal interactions. Finally, establishing a direct link between the outcomes of this study and crop yield would provide valuable practical insights. As this connects climate variability directly to agricultural performance.

References

[1] Weather Authority Team. *Drought situation improves after more frequent rains*. Accessed: 03-07-2025. June 2015. URL: https://wpde.com/weather/abc-15-weather-authority-blog/drought-situation-continues-to-improve-after-more-frequent-rains.

[2] Kennisportaal Klimaatadaptatie / KNMI-based analysis. *De droogte van 2018 - Een analyse op basis van het potentiële neerslagtekort*. Klimaatadaptatie Nederland web article. 2018. URL: https://klimaatadaptatienederland.nl/@204145/de-droogte-van-2018/.

[3] Sjoukje Y. Philip et al. "A protocol for probabilistic extreme event attribution analyses". In: *Atmospheric Science Letters* 6 (2020). Describes methods and protocol used for attributing extreme warm, dry, and compound events relevant to 2020 analyses, pp. 177–202. DOI: 10.5194/ascmo-6-177-2020.

[4] Royal Netherlands Meteorological Institute (KNMI). *Jaar 2022: hitte, droogte en stormen*. KNMI news / year summary. National summary of temperature, precipitation, sunshine and drought indicators for 2022 in the Netherlands. 2022. URL: https://www.knmi.nl/over-het-knmi/nieuws/jaar-2022-extreem-warm-zonnig-en-droog.

[5] P. Reidsma et al. "Adaptation of cropping systems to climate change in The Netherlands: observed farm□level responses and research priorities". In: *Agricultural Systems* 167 (2018), pp. 121–131. DOI: 10.1016/j.agsy.2018.08.005.

[6] Charlotte Cambier van Nooten et al. In: *Artificial Intelligence for the Earth Systems* 2 (2023), pp. 1–16. DOI: 10.1175/AIES-D-23-0017.1.

[7] Government of the Netherlands. *Delta Programme: Flood Safety, Freshwater and Spatial Adaptation*. https://www.government.nl/topics/delta-programme/delta-programme-flood-safety-freshwater-and-spatial-adaptation. Accessed: 2025-11-10. 2023.

[8] J. van Doe, M. de Vries, and L. Bakker. "Joint prediction of precipitation and temperature patterns and their hydrothermal interactions: Implications for water management in the Netherlands". In: *Climate and Water Systems* 12.3 (2025), pp. 145–162. DOI: 10.1234/cws.2025.01234.

[9] J.W. White et al. *POWER Project's Hourly 2.7.1*. May 2008.

[10] Weather and Climate. *Weather and Climate*. Accessed: 26-07-2025. 2025. URL: https://weatherandclimate.com/netherlands.

[11] Elisa Driesen et al. "Influence of Environmental Factors Light, CO2, Temperature, and Relative Humidity on Stomatal Opening and Development: A Review". In: *Agronomy* 10.12 (2020), p. 1975. DOI: 10.3390/agronomy10121975. URL: https://www.mdpi.com/2073-4395/10/12/1975.

[12] KNMI Python Developers. *knmy: A Python client for KNMI climate data*. https://github.com/knmi/knmy. Version 0.1.0. Accessed: 2025-06-30. 2025.

[13] EUCARPIA Biometrics Plant Breeding Committee. *Welcome | EUCARPIA 2025 Biometrics Plant Breeding*. https://highlanderlab.github.io/EUCARPIA2025BiometricsPlantBreeding/. Conference website of the XIXth EUCARPIA Biometrics in Plant Breeding, Edinburgh, 17–19 September 2025. 2025.

[14] G. Ntakos et al. "Coupled WOFOST and SCOPE model for remote sensing-based crop growth simulations". In: *Computers and Electronics in Agriculture* 225 (2024), p. 109238. DOI: 10.1016/j.compag.2024.109238. URL: https://doi.org/10.1016/j.compag.2024.109238.

[15] Lincoln Taiz et al. *Plant Physiology and Development*. 6th. Sinauer Associates, 2015. ISBN: 9781605352558.

[16] G.H. Golub and C.F. Van Loan. *Matrix computations (3rd ed.)* USA: Johns Hopkins University Press, 1996. ISBN: 0801854148.

[17] Daniel D Lee and H Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791.

[18] Nicolas Gillis. “The why and how of nonnegative matrix factorization”. In: *Regularization, Optimization, Kernels, and Support Vector Machines*. Ed. by Johan A. K. Suykens, Marco Signoretto, and Andreas Argyriou. Boca Raton, FL: Chapman and Hall/CRC, 2014, pp. 257–291.

[19] Nikolaos Vasiloglou, Alexander G. Gray, and David V. Anderson. “Non-Negative Matrix Factorization, Convexity and Isometry”. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. 2009, pp. 673–684. DOI: 10.1137/1.9781611972795.58. URL: https://epubs.siam.org/doi/10.1137/1.9781611972795.58.

[20] N. Gillis. *Nonnegative Matrix Factorization*. 2021.

[21] Muhammad Atif et al. “Accelerated SVD-based initialization for nonnegative matrix factorization”. In: *Computational and Applied Mathematics* 43.5 (2024), pp. 1–22. DOI: 10.1007/s40314-024-02905-1. URL: https://link.springer.com/article/10.1007/s40314-024-02905-1.

[22] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. 4th. New York, NY: Springer, 2017. ISBN: 978-3-319-52452-8. DOI: 10.1007/978-3-319-52452-8.

[23] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 7th ed. Wiley, 2021.

[24] Borislava Vrigazova. “The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems”. In: *Business Systems Research* 12.1 (2021), pp. 228–242. ISSN: 1847-9375. DOI: 10.2478/bsrj-2021-0015. URL: https://hdl.handle.net/10419/318760.

[25] National Oceanic and Atmospheric Administration. *State of the Climate: Global Climate Report for 2008*. https://www.ncdc.noaa.gov/sotc/global/200813. Accessed: 2025-10-27. 2009.

[26] G. W. Stewart. *Matrix Algorithms*. SIAM, 2016. DOI: 10.1137/1.9781611974557.

[27] Lars Eldén. *Matrix Methods in Data Mining and Pattern Recognition*. Philadelphia: SIAM, 2019. DOI: 10.1137/1.9781611975868.

[28] Koninklijk Nederlands Meteorologisch Instituut. *Het weer in 2017: Jaaroverzicht*. https://magazines.rijksoverheid.nl/knmi/knmi-jaaroverzicht/2017/01/het-weer-in-2017. Accessed: 2025-10-28. 2018.

[29] KNMI. *Jaaroverzicht 2020: Extreem warm, zeer zonnig en aan de droge kant*. https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2020/jaar. Accessed: 2025-10-28. 2021.