GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Explaining Agricultural Productivity using Meteorological Indicators: an Exploratory and Predictive Analysis

Marisa Lange [1], Laura Plodek [2,*]

[1] *Department of Economics, Georg-August-University of Göttingen, Germany*
[2] *Department of Economics, Georg-August-University of Göttingen, Germany*

Abstract Agricultural yield modeling plays a crucial role in facilitating informed decision-making for farmers and stakeholders in the agricultural sector. However, the complexity of environmental factors and the presence of multicollinearity among predictor variables pose challenges to traditional modeling approaches. In this study, we investigate the effectiveness of advanced statistical methods and innovative feature engineering strategies in addressing these challenges on the example of Canola production in the province of Saskatchewan, Canada. We propose the computation of average maximum temperatures and standard precipitation indices during critical months of the crop life cycle as model features. We contrast the use of conventional linear regression models with machine learning approaches with particular regard to their ability to handle multicollinearity. We put our findings to use and approach predictive modelling in two different manners, drawing conclusions from the respective performances. Our findings contribute to the advancement of agricultural yield modeling by providing a comprehensive framework that integrates advanced statistical techniques, innovative feature engineering strategies, and a standardized data processing pipeline.

Keywords Total Variance Explained, Precision Agriculture, Yield Prediction, Feature Engineering, Decision Support, Multicollinearity, Spatial Dependencies, Temporal Dependencies, Model Explainability

*Correspondence to: Marisa Lange (Email: marisa.lange@stud.uni-goettingen.de). Department of Economics, Georg-August-University of Göttingen. Platz der Göttinger Sieben 3, 37073 Göttingen, Germany.

1. Introduction

Little explanation is needed on the crucial role that agriculture plays in ensuring food security and sustaining global economies. Demand for primary crops, particularly for the purpose of consumption, is showing consistent increase from year to year (+2% from 2020 to 2021 [29]) amid the lasting global growth in population [52]. More than ever, the sector faces the challenge of having to increase supply efficiency. This is primarily attributable to resources becoming more scarce and climatic conditions decreasing in predictability and suitability amid a rise in extreme temperature occurrences [14]. Political pressure to comply with increasingly strict policies, e.g. to reduce fertilizer emissions [55], only adds to the dilemma many farmers and participants in the agricultural market face. Resolving these conflicts is of utmost importance as global food security has to be ensured and made sustainable.

All of these developments point towards the increased importance of sustainable production planning and proactive adaptation to ever-changing external conditions, a hurdle to the sector that continues to heavily rely on traditional practices [40]. A promising approach that has arisen over the recent years and continues to evolve through new developments is the enhancement of agricultural practices by applying advanced technology and data analytics solution, widely known as precision agriculture [10]. Driven by innovative technology start-ups and governmental aid alike [2], precision agriculture practices show potential in various growing-related activities, such as the use in fields and activity planning. When used in the production of large-scale crops, they can lead to significant and lasting impact on the economy as a whole. One of these major global crops is canola, also known as rapeseed [5].

For precision agriculture modelling to maintain validity over long-term periods and ever-changing conditions, a sufficient degree of domain knowledge is required in order to define sensible input factors [46]. Key external factors influencing primary crop yield have generally proven to be the temperature as well as precipitation amounts [34]. What is yet to be explored further is the specific nature of the type of relation these variables have to the relative yield, as well as the relation in-between the variables. Questions that remain include whether influences are better captured on the basis of longer-term averages, or if extreme events play a more crucial role in the influencing growing conditions. Analogously, spatial fluctuations on a local scale have yet to be explored in order determine how scaled down these meteorological indicators have to be in order to be suitable as model input. Finally, while temperature and precipitation undoubtedly play major roles in shaping growing conditions, they are likely not to be the only determinants. Importance of other variables, such as the type of soil the crop is planted on, has yet to be investigated with particular regard to its relation to climatic factors.

The purpose of the work done in this paper is to explore the ways in which climatic and other potentially crop-influencing variables impact Canola yield in one of the major production areas, the region of Saskatchewan in Canada. We investigate relations between these variables and the problems they may cause in predictive modelling applications. Attempting to capture relevant information efficiently and accurately, we propose the derivation of standardized features and investigate further processing techniques, which we evaluate through various linear and non-linear predictive modelling approaches. As the development of precision agriculture moves forward and certain practices begin to emerge as industry standards across crops and regions [38], we aim to demonstrate the persisting importance of taking domain-specific particularities into account. To avoid over-tailoring this solution to one crop and accounting for scalability of our approach to various crops and geographical areas, we propose a universal feature engineering and modelling approach that is robust to condition changes and, for the purpose of explainability, allows for insights into individual feature importance.

2. Theoretical Background

Reports for 2021 projected the global agricultural market grow by around 10.7 percent to over 18 billion USD in 2026 [28]. It includes animal and crop production as well as the sale of related agricultural services (e.g., equipment provision or chemical and fertilizer retailing). Having grown by more than 50% since the beginning of the century [29], production of primary crops totaled 9.5 billion USD in 2021 and accounts for around 40% of the total agricultural market volume.

Aside from cereals, sugar crops and vegetables, oilseed crops account for 12% of primary crop production and thus make up one of the major categories when segmenting this market by commodity. In terms of worldwide production volume of oilseeds, canola is the second biggest produce in the category, falling short only to soybeans with a forecasted 87.44 million tonnes to be produced during the 2023/24 period [48]. A versatile crop, canola is grown as a basis for edible oil as well as for numerous industrial applications. These range from biofuel to plastics and cosmetics production as well as livestock feed [24]. Canola is derived from seeds of the rapeseed plant and crushed to oil and meal. While rather sensitive to soil acidity, the plant is known to be relatively robust to the texture of the ground it grows on. There are both winter and spring varieties of canola plants, although the main producing regions tend to focus on the latter [26]. The cultivation stages of spring varieties can be roughly outlined as a 16 to 18-day germination period, followed by a 30-day vegetative period. After an up to 20-day flowering period, the plant has a ripening period of around 20 days [26].

Ahead of germinating, plants require moist seedbeds at temperatures of at least 2, but ideally 10 degrees Celsius or higher. Once at the flowering stage, however, temperatures surpassing 30 degrees Celsius have a detrimental effect.

2.1. Canola Cultivation in Saskatchewan

A crop native to Canada, the country remains the world leader in canola production. With a national production volume of over 13 million tonnes in 2021 [50]. Canola is grown on around 21.5 million acres of land in Canada. Fields are predominantly located in the Southwestern regions of the country, which have proven to provide the best suited external conditions with regard to temperature and daylight requirements, precipitation patterns, and soil quality [43].

With wide areas of flat land, the Prairie province of Saskatchewan has almost 40 percent and thus the leading share of total national farm area [8]. It moreover accounts for half of Canada's crop production, providing employment to over 55 thousand individuals [13]. Agriculture makes up around 11 percent of the region's GDP and accounts for just over 5 percent of total employment [31].

While total acreage for oilseed crops has remained relatively stable since 2016, yield in bushels per acre has shown significant increase for canola in Saskatchewan and displayed total volume growth of 65.47% from 2000 to 2020 [23]. This increase in efficiency can largely be contributed to technological advancements in farming techniques (e.g., operational management and irrigation planning). While undeniably playing an important role in shaping the annual output increases over the past century, these effects remain diffcult to quantify [4]. Examples in Saskatchewan include the use of auto-steer equipment, drones as well as Geographic Information System (GIS) mapping.

Canola farmland in Saskatchewan is distributed over three predominant soil zones. While the majority of canola is grown on black and dark soil, expansions into the brown soil zones are ongoing. These developments can be attributed to growing demand as well as the development of more heat and drought tolerant canola varieties [51].

2.2. Statistical Methods

In this section, we offer a concise overview of the statistical methodologies employed throughout the paper. This summary includes the techniques and analytical strategies utilized to address our research, highlighting specific statistical tests, models, or computational methods applied to the data. The focus is on providing a clear understanding of how these methodologies contribute to the findings and conclusions of the project.

2.2.1. Analytical Methods

Standardized Precipitation Index The Standardized Precipitation Index (SPI) is a widely used indicator in meteorology and climate research for identifying changes in precipitation amounts for a specific time period and location, relative to the historical precipitation observed for that location and period. This makes the SPI an apt indicator for identifying droughts.

The calculation and interpretation of the SPI proceed as follows: Initially, one must define the location and the time period of interest, with the time period usually ranging between 1 and 12 months. Utilizing precipitation data for the specified period and location, a distribution that fits the historical record is determined. The gamma distribution is a common choice for this purpose

$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)},$$

where $x > 0$, $k > 0$ is the shape parameter, $\theta > 0$ is the scale parameter, and $\Gamma(k)$ is the Gamma function evaluated at k. The scale and shape parameters of the gamma distribution are then fitted to the accumulated precipitation in the historical data, using approximations of maximum likelihood estimators.

For any non-zero amount of precipitation, a cumulative probability is derived. With accounting for the possibility of no precipitation, the cumulative probability of the observed rainfall is then transformed into a standard normal variable with a mean of zero and a variance of one. Consequently, the SPI is expressed in units of standard deviation from the mean of the standardized distribution, based on historical data.

The SPI can be interpreted in a manner akin to a normal standard deviation. Values ranging between -1 and 1 are considered to indicate normal precipitation levels. Values between -1 and -1.5 or 1 and 1.5 are deemed as moderately wet or dry, respectively, while values up to -2 or 2 are viewed as very dry or wet. SPI values exceeding -2 or 2 indicate extreme conditions [47].

Moran's I Moran's I is a statistical index used to identify and quantify spatial autocorrelation. Spatial autocorrelation occurs when objects that are closer together are more similar than those further apart, leading to the formation of clusters. The index is calculated by

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

where N is the number of spatial units indexed by i and j, x_i and x_j are the observations at locations i and j, $\bar{x}$ is the mean of the observations, w_{ij} is a spatial weight between locations i and j, and W is the sum of all w_{ij}.

Moran's I values range from -1 (perfect negative autocorrelation) through 0 (randomness) to +1 (perfect autocorrelation). To calculate spatial weights, neighborhood structures, like the K-nearest neighbors algorithm, are can be utilized [39].

Variogram The variogram is a key tool for analyzing spatial dependency structures. It quantifies the correlation strength between spatial observations by measuring differences based on distance, typically in lags. The empirical variogram is estimated using the semivariance through the formula:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left(z(x_i) - z(x_i + h)\right)^2 \tag{1}$$

where $\hat{\gamma}(h)$ is the estimated variogram for a lag distance h, $N(h)$ is the number of data pairs separated by the distance h, and $z(x_i)$ and $z(x_i + h)$ are the values of the variable at locations x_i and $x_i + h$, respectively.

The variogram's outcome provides semivariance values at specific lags. This foundation allows for fitting a theoretical variogram by selecting an appropriate correlation function, such as Gaussian or exponential. The theoretical variogram, a continuous function, facilitates semivariance estimation across any distance, enhancing the analysis of spatial relationships [54].

Interpolation with Radial Basis Functions Radial basis functions are functions of the euclidean distance between two locations, i.e. $B(\|z - k\|)$, where z and k are the locations and the function B is the radial basis function, which can take different shapes. Common choices for interpolation tasks are e.g. the Gaussian radial

basis function $g_j = e^{-cr_{ij}^2}$ or the thin plate spline radial basis function $g_j = r_{ij}^{2m}\log(r_{ij})$, where r is the euclidean distance between locations i and j. For Thin Plate Spline the interpolation can be expressed as:

$$f(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \sum_{j=1}^{n} \gamma_j B_j(z_1, z_2),$$

where B_j are the thin plate spline functions and $\beta_0, \beta_1, \beta_2$ are linear coefficients and γ_j are coefficients that work as weights [27].

Feature Correlations Correlation coefficients quantify the strength and direction of the linear relationship between two correspoding features. In a set of multiple features, they are obtained by means of calculating the correlation matrix.

$$\rho_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}} \tag{2}$$

Where ρ_{ij} is the correlation coefficient between features i and j, x_{ik} and x_{jk} are the values of features i and j for observation k, $\bar{x}_i$ and $\bar{x}_j$ are the means of features i and j respectively, and n is the number of observations.

Principle Component Analysis Principal Component Analysis (PCA) is a widely-used multivariate technique aimed at reducing dimensionality, particularly beneficial for datasets afflicted by multicollinearity [32]. Multicollinearity, characterized by high correlations among variables, signifies the existence of constant or near-constant linear relationships among them. This phenomenon can result in unstable coefficient estimates in regression models and diminish the interpretability of individual predictors. PCA operates on the premise of reducing the dimensionality of a dataset by constructing linear combination vectors of its individual variables, thereby addressing correlations between them. Through this process, the majority of information from the original variable set is preserved. Each derived component captures a distinct proportion of the total variance present in the data, with components organized based on their respective explained variability. Resulting uncorrelated principal components facilitate the application of regression techniques.

PCA relies on principles derived from matrix theory. Initially, the covariance matrix of the dataset is computed, followed by the determination of its eigenvectors and eigenvalues, which are utilized to identify the principal components.

$$\text{Covariance Matrix:} \quad \Sigma = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}) \tag{3}$$

$$\text{Eigendecomposition:} \quad \Sigma v = \lambda v \tag{4}$$

$$\text{Principal Components:} \quad Z = XV \tag{5}$$

The symmetric covariance matrix Σ computed in 3 measures the relationships between pairs of variables in a dataset. X represents the data matrix, and $\bar{X}$ denotes the mean vector of the data computed across each variable. Σ is computed through taking the average of the outer product of the centered data matrix $(X - \bar{X})$ with itself.

Through the Eigendecomposition outlined in 4, Σ is then decomposed into its constituent eigenvectors and eigenvalues. v represents the eigenvectors of the covariance matrix, and λ the corresponding eigenvalues. Multiplying Σ by an eigenvector results in a scaled version of the eigenvector, where the scaling factor is the corresponding eigenvalue.

Principal Components (PCs) are finally obtained by projecting the original data onto the eigenvectors of the covariance matrix as in 5. Z represents the transformed data matrix, where each column corresponds to a principal component. X is the original data matrix, and V contains the eigenvectors of the covariance matrix Σ [1].

2.2.2. Regressive Methods In an attempt to overcome the multicollinearity issue implicitly, we fit several machine learning models known to be less susceptible to issues arising from this. [20]

LASSO Regression LASSO, or Least Absolute Shrinkage and Selection Operator, is a regularization technique enhancing least squares estimations, particularly for complex, high-dimensional data prone to multicollinearity. It introduces a penalty for complexity governed by the hyperparameter λ, optimally determined through a cross-validation. LASSO's penalty term is the sum of the absolute values of the coefficients, leading to:

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \right\}, \tag{6}$$

where y is the response vector, X is the matrix of predictors, β is the vector of coefficients, k is the number of predictors, and λ is the regularization parameter [27].

The penalty mechanism in LASSO aims to reduce the values of smaller coefficients to zero, effectively removing them from the model, while larger coefficients remain less affected. This characteristic not only helps in managing multicollinearity but also enables LASSO to serve as a method for variable selection, simplifying the model by identifying and retaining only the most significant predictors.

Ridge Regression Ridge Regression, akin to LASSO, introduces a penalization approach. The difference between the two methods lies in the definition of the penalty term. For ridge the penalty term is intriduced as the sum of squared coefficients. This causes coefficients to shrink but not to zero, preserving all variables in the model. Thus, Ridge is particularly useful for mitigating multicollinearity without the goal of variable selection, making it suitable for cases where retaining all predictors is desired [27].

Support Vector Regression Support Vector Regression (SVR), a supervised approach derived from statistical learning theory [53], extends the concept of support vector machines (SVMs) to regression problems. It utilizes a non-linear mapping technique to project input vectors into a higher-dimensional feature space, where a separating hyperplane is sought. This hyperplane, combined with an appropriate learning algorithm, aims to capture the relationship between input features and the target variable [41].

The SVR model is represented in 7, where $f(x)$ predicts the target variable, α_i are the coefficients determined during training, $K(x_i, x)$ is the kernel function measuring the similarity between input vectors, and b is the bias term.

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{7}$$

SVR differs from traditional linear regression by minimizing the margin of error within a specified threshold around the predicted values. The width of this margin is controlled by the parameter ϵ. Lower values of ϵ result in a narrower margin, fitting the training data more closely, while moderate to large values lead to higher generalization.

The kernel function K handles the transformation of the input features. We adopt the Radial Basis Function (RBF), i.e. the Gaussian kernel, which is commonly used to capture potential non-linearities in data.

$$K(x_i, x) = \exp\left(-\gamma \cdot |x_i - x|^2\right) \tag{8}$$

In 8, γ specifies the kernel coefficient, taking influence on the shape of the decision boundary. We set γ using the *'scale'* parameter, calculated as $\frac{1}{n_{\text{features}} \cdot X.\text{var}()}$, where n_{features} is the number of features in the dataset.

Support Vector Machines (SVMs) further include a regularization parameter C, controlling the trade-off between the training error and model complexity. An l2 penalty, or Ridge regularization, is applied. Smaller values of C prioritize generalization power, while larger values result in a closer fit to the training data [45].

Random Forest Regression Random Forest Regression is a non-parametric ensemble method used for regression tasks, built upon decision trees [36, 9]. Random Forests utilize multiple decision trees and train each one on a random subset of the data, subsequently controlling over-fitting through averaging [45]. Through combining the predictions of individual trees, random forests form an additive model, represented by 9.

$$g(x) = \sum_{i=0}^{n} f_i(x) \tag{9}$$

$g(x)$ is the predicted target variable for the input vector x based on the random forest regression model, computed as the aggregate of predictions made by individual trees in the ensemble $f_i(x)$.

The importance of features in random forests is assessed using Gini impurity. For each feature, its contribution to the overall reduction in impurity across all trees is calculated. Features with higher importance scores indicate greater predictive power.

Serving the purpose of tuning the model to balance complexity with generalization power, key parameters in practice include the following [45]:

- **n_estimators**: The number of decision trees in the ensemble.
- **criterion**: The quality measurement function for splits; typically MSE or MAE.
- **max_depth**: The maximum depth of each decision tree.
- **min_samples_split**: The minimum number of samples required to split an internal node.
- **min_samples_leaf**: The minimum number of samples required at a leaf node.
- **max_features**: The maximum number of features considered for splitting.

Long Short-Term Memory Network (LSTM) Long Short-Term Memory (LSTM) networks represent a class of gradient-based neural network models particularly well-suited for handling sequential or time series data. A subtype of recurrent neural networks (RNNs), LSTMs are designed to address the challenges of learning long-term dependencies in sequential data while mitigating the vanishing gradient problem [37].

A network's architectural structure comprises several key elements, including memory cells, activation functions, hidden layers, and delays. Memory cells store information over time and thereby enable the network to retain long-term dependencies. Activation functions steer the flow of information through each unit within a network. Hidden layers, containing interconnected memory cells, allow the network to capture complex patterns and relationships within the data. Delays incorporate time connections between memory cells and facilitate the processing of input sequences in a sequential manner and thus capture temporal dependencies [49].

Training an LSTM network involves optimizing parameters through the back-propagation algorithm, which computes gradients of the loss function with respect to the network parameters. Key adjustment options to optimize the performance of an LSTM include the learning rate which determines the step size of parameter updates during training, the batch size specifying the number of input samples processed in each training iteration, as well as the number of training epochs.

For a comprehensive understanding of LSTM networks and their optimization, further elaboration is available in the seminal publication by Hochreiter and Schmidhuber [30].

3. Data

3.1. Annual Canola Yield in Saskatchewan

The Saskatchewan government provides data on agricultural yields for various crops across many regions (crop districts) in the agricultural region of Saskatchewan, available from different years, depending on when each crop was first cultivated [42]. For Canola, data continues to be published annually since 1971 and is accessible in various formats. The yield is measured in bushels per acre, a unit common for agricultural yield in the US and Canada. This unit was originally a measure of volume and is used today as a unit of mass varying by crop. In the context of canola, one bushel equals approximately 50 pounds (22.68 kg).

3.2. Temperature and Precipitation in Saskatchewan

The Copernicus Climate Change Service (C3S), as a part of the European Union's Copernicus Earth observation program, has established a comprehensive global weather database. This publicly accessible database is designed

to offer detailed, authoritative information on climate change and its impacts and providing crucial insights into climate dynamics [22].

From this resource, we obtained specific weather data for each crop district in Saskatchewan, covering every year from 1971 to 2022. The dataset includes hourly measurements throughout Canola's growing season, from April to October, featuring temperatures in Kelvin and precipitation amounts in meters. This extensive dataset provides a valuable foundation for analyzing weather patterns and their impacts on agricultural productivity within the region.

3.3. Soil Types in Saskatchewan

Information on the dominant soil type for each municipality in the province is provided through an interactive map by the Saskatchewan Crop Insurance Corporation (SCIC) [44].

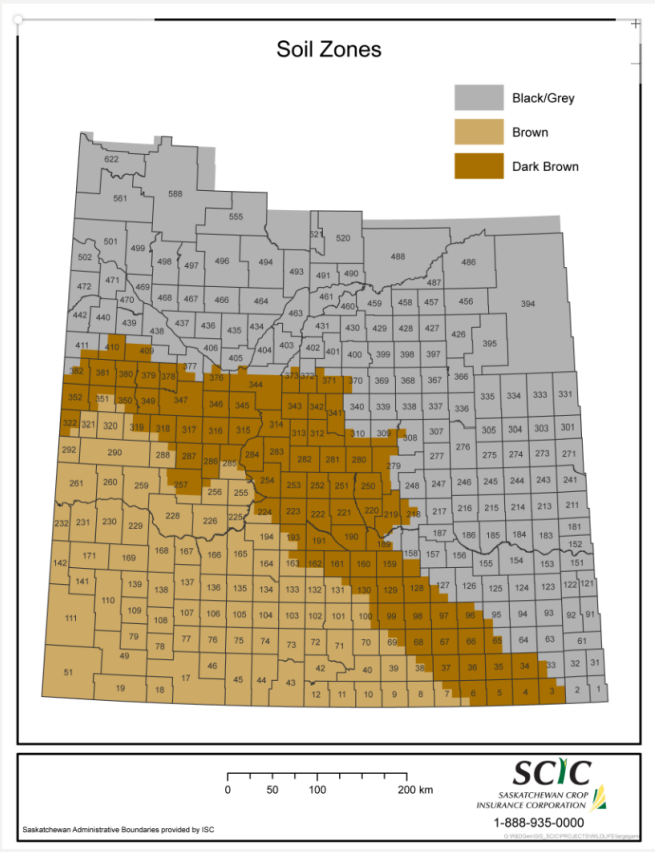


Figure 1. Soil Type by Administrative Boundary in the region of Saskatchewan, Canada [44].

3.4. Processing

3.4.1. Canola yield After obtaining the raw data, the initial step involved its reduction to a sufficiently large but complete subset of information that could be analyzed and interpreted sensibly as well as predicted. As the data spanned across a 50-year time horizon as well as a total of 622 individual geographical regions, temporal and spatial completeness had to be accounted for. We observed a large number of missing values particularly within the first 20 years of data, i.e. between 1970 and 1989. In accordance with domain-specific characteristics, these years are subsequently excluded in this work, as the disadvantages of the incompleteness outweigh the potential information gain from including this time period.

Following the reduction of the time horizon, a total of 181 regions with full data availability remained. These center largely around the Southeastern to Central part of the province, which are known to be the main locations of Canola growing.

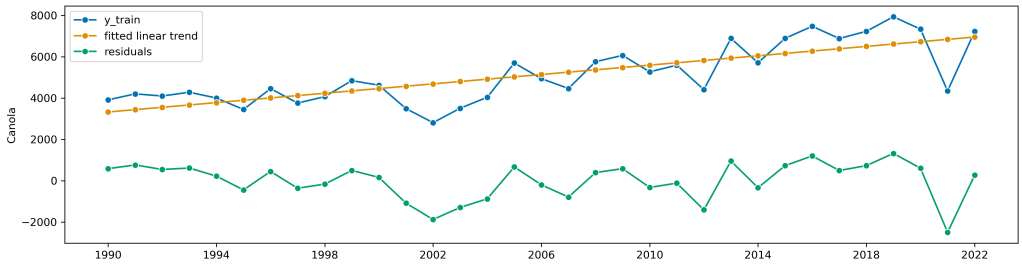


Figure 2. Process of de-trending on the time series of the cumulative yield over all regions.

Furthermore, Within the annual Canola yield values, there was a notable long-term growth trend pattern in the latter observations. This observed trend in increasing Canola yields over the years can largely be attributed to recent technological advancements in agriculture. However, experts anticipate this trend will not continue due to two main reasons: firstly, the pace of technological improvements is expected to slow amid diminishing returns, as it becomes increasingly challenging to achieve significant breakthroughs or enhancements. Many fields are reaching a point where the most significant gains have already been made, and further improvements require increasingly substantial investments. Simultaneously, the impact of the climate crisis on agricultural productivity presents significant challenges that may offset any gains from technological progress, potentially leading to stagnation or even declines in food production in some regions [21].

In this study, we aim to delve deeper into the impact of extreme weather events on Canola yield, observing that these events manifest as local fluctuations and peaks within the yield time series. The overarching trend, while notable, is considered less critical to our analysis and rather difficult to capture accurately [4]. Consequently, we opted to linearly de-trend the time series, focusing instead on the residuals to better isolate and examine the effects of weather extremes. This results in the time series being centered around zero. The de-trending process is illustrated in 2, enabling a clearer analysis of the relationship between weather events and yield variability.

Given that yields are quantified on a per-acre basis, adjusting for overall yield discrepancies across different districts isn't necessary. This standard measurement allows for direct comparison of productivity irrespective of the total land area cultivated in each district.

3.4.2. Weather data As the 181 regions complete in Canola yield data sufficiently represented the area of interest, we excluded all remaining incomplete regions and moved forward using the following information basis:

Xarray arrays are built on top of NumPy arrays and include labeled dimensions, making it easier to understand multi-dimensional data in principle. They extend the functionality of NumPy arrays by adding labeled dimensions, coordinates, and metadata [56]. Although powerful, the complexity of the format proved to be rather inconvenient prior to our analysis, and we therefore reduced the data to a conventional a two-dimensional dataframe. This was done using geographical centerpoints by latitude-longitude pairs to assign IDs to each region and store all measurements with a column *'region'*, preserving spatial information. *'Time'* was preserved as a column as well.

A key question was how to align the availability of the climate variables *'t2m'* and *'tp'* in hourly measurements with the Canola yield, for which we only had annual amounts. We decided to denote each timestamp with the de-trended annual Canola yield value corresponding to its year, which would later allow us to aggregate time periods as sensible. An additional factor variable *soil_zone* was computed manually based on [44], where each of region was assigned a number for its predominant soil type.

The intermediate dataframe thus consisted of the six columns as detailed in 2.

Type	Name	Description	Details
Dimensions	longitude latitude time		88 unique values 41 unique values 168.488 unique values
Coordinates	longitude latitude time		[-110.0, -101.3] [53.0, 49.0] [1990-04-01T00:00:00, 2022-10-31T23:00:00]
Data Variables	t2m	air temperature at 2 metre height above surface [7]	(5136, 41, 88)
	tp	total amount of condensation of atmospheric water vapor falling from clouds due to gravitational pull in the specified period [3]	(5136, 41, 88)

Table 1. Summary of the xarray dataset containing multi-dimensional 2-metre temperature and total precipitation values.

Time	t2m	tp	canola_detrended	region	soil_zone
datetime	float64	float64	float64	int	factor

Table 2. Intermediate Raw Information Dataframe.

3.5. Feature Engineering

A key aspect of this project involved constructing significant features from weather data, serving as the foundation for our analysis. This task necessitates multiple considerations: identifying appropriate features that reflect domain knowledge, ensuring the interpretability of results derived from these features, and maintaining an optimal balance between maximizing the dataset's informational value and preventing multicollinearity arising from overly similar features.

In our analysis, aiming to gain deeper insights into the impact of extreme weather events on the Canola yield, it is essential to concentrate on features that specifically address occurrences such as unusually high temperatures or prolonged dry periods. The frequency of such weather events has increased in recent years, a trend further detailed in section 3.7. Given the rapidly advancing climate crisis, understanding these events will become increasingly critical for agriculture in the near future.

There exists a substantial theoretical foundation regarding which weather occurrences are beneficial or tolerable for the Canola plant and which events pose the most significant challenges. Further details on these assumptions are provided in the introductory chapter. Incorporating this prior knowledge is crucial when defining features.

The weather data available to us encompass temperature and precipitation measurements taken hourly throughout the Canola growing season, which spans from April—when Canola begins to grow—to October, when it is harvested. Given that the target variable of our analysis, the Canola yield, is assessed annually, it is necessary to identify features that align with this observation period. Consequently, despite possessing hourly measured data, we must aggregate this information to derive features that reflect an annual perspective as well.

In our analysis, we aim to concentrate on three key aspects for feature selection. Firstly, we will focus on the effects that are known to be the most crucial for Canola, while disregarding those not deemed problematic for this crop specifically. Secondly, we will place a stronger emphasis on the timing of weather events throughout the year. As detailed in 1, Canola undergoes several developmental stages over the growing season. For instance, there is

a theoretical basis for the assumption that the plant is particularly vulnerable during the reproductive phase. This implies that a drought or heatwave during this phase could have more severe consequences than the same event occurring at a different time. Thirdly, we aim to highlight the occurrence of extreme weather events. To achieve this, we analyzed a historical time span and compared the weather conditions of the period under study to this previous time frame. Given our assumption of an increase in extreme weather events in recent years, examining past years can aid us in identifying these unusual weather occurrences, e.g. with the use of quantiles.

In the subsequent subsections, a detailed description of the features utilized in our analysis is provided. Furthermore, a summary of these features is available in 3. The Code on how we derived the features an be found in the appendix A.

3.5.1. Monthly weather features For this category of features, capturing the general behavior of the weather within a single month was crucial. We aimed to capture both the overall level of temperature or precipitation for the month and the frequency of notably high or low occurrences.

For temperature, we choose to calculate the monthly average of the daily maximum temperatures. This approach allows us to fulfill both criteria: obtaining a general level for the month while also accounting for temperature peaks. We chose not to consider the lowest temperature points in this variable, based on theoretical background indicating that Canola is quite resilient to unusually low temperatures, especially during the summer months. The potential impact of frost days on the plant is addressed separately in the extreme weather indices. Using a single variable to describe each month's temperature offers a significant advantage over introducing several variables for each moth (e.g., one for the average and one for the maximum temperatures or the number of days exceeding a certain temperature threshold), which would inherently be highly correlated. By employing just one variable, we effectively minimize the risk of introducing additional sources of collinearity into our features.

For precipitation, we aimed for a similar approach, focusing on two key factors: the total amount of precipitation within a month and the distribution of this precipitation. The impact of the same volume of rainfall can vary significantly, depending on whether it occurs over a few days or is evenly spread throughout the month. Capturing the distribution of rainfall with monthly features presents challenges, with various considerations discussed further at the chapter's end and in the section dedicated to extreme weather indices.

For our monthly precipitation variable, we opted to use the Standardized Precipitation Index (SPI). This index compares the observed amount of precipitation over a specific period, in our case the month, to the historical precipitation distribution for that same period. A comprehensive explanation of how the SPI is calculated is provided in section 2.2. Employing the SPI allows us to quantify how the precipitation for each month stands relative to what is historically typical, offering a direct measure of whether the month was unusually dry or wet.

3.5.2. Extreme weather indices In addition to the monthly features, we have incorporated indices for extreme weather events. These indices are measured across the entire growing season, which means we do not obtain information on how the timing of an event—occurring in different months or developmental stages of the plant—affects the plant. We excluded this aspect because, for certain extreme weather conditions, their occurrence is tied to specific times of the month. For instance, frost days are predominantly observed in October and April, while problematic heat days are more common during the summer months. Other indices are designed to consider the entire growing period, and calculating them on a monthly basis would undermine their purpose. For example, tracking consecutive days of an event within a single month could potentially interrupt a sequence that spans from the end of one month to the beginning of the next. The extreme weather indices we utilize are identical to those applied by [35]. These indices were defined by the CCI/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices.

The indices can be categorized into two distinct types. The first group includes the counting of days characterized by very low and high temperatures, as well as days with an exceptionally high amount of precipitation. For temperature, frost days are counted, defined as days when the minimum temperature falls below 0 degrees Celsius. Summer days are those where the maximum temperature exceeds 25 degrees Celsius. To identify extremely wet days, we utilized the 95 % quantile for daily precipitation during the reference period from 1971 to 1989. A day is considered extremely wet if its precipitation surpasses this threshold. We focus exclusively on the upper quantile

Variable Type	Variable Name	Description	Unit
Monthly-Weather Indices	Avg-Max-Temp in Month	The average of the daily maximum temperatures in one month, for the Month April to October.	Kelvin (K)
	SPI in Month	The Standardized Precipitation Index for one month, for the Month April to October.	Standard Deviations
Extreme-Weather Indices	Summer Days	Days during the growing season when the max. temperature exceeds 25 $°C$.	Number of Days
	Frost Days	Days during the growing season when the minimum temperature falls below 0 $°C$.	
	Longest Dry Spell	Maximum number of consecutive days during the growing season with precipitation lower than 1 mm.	Number of Days
	Longest Wet Spell	Maximum number of consecutive days during the growing season with precipitation larger than 1 mm.	Number of Days
	Extremely Wet Days	Days during the growing season when precipitation is greater than the 95th quantile of the reference period	Number of Days.
	Longest Heat Wave	Maximum number of consecutive days during the growing season when the maximum temperature exceeds the 90th quantile for the reference period.	Number of Days
	Longest Cold Wave	Maximum number of consecutive days during the growing season when the minimum temperature falls below the 10th quantile for the reference period.	Number of Days.
Non-Weather Variables	Soil Zone	Category variable, indicating into which soil zone the observation falls.	Categories $\{1,2,3\}$

Table 3. Summary of the explanatory variables used for analysing the impact of (extreme) weather events on the canola yield.

for precipitation since days without any rainfall are obviously very common. Droughts over a longer period of time, on the other hand, are better represented by consecutive dry days, which are addressed by subsequent indices.

The second category focuses on the longest sequence of consecutive days exhibiting specific unusual weather patterns, considering both temperature and precipitation. For precipitation, we examine periods both with and without rain, ultimately identifying the longest stretch within the entire growing period that experienced daily rainfall and the longest period of complete dryness.

Regarding temperature, we compare daily maximum and minimum temperatures to the reference period of 1971-1990. We count the days on which the maximum or minimum temperatures surpass the 90th percentile or, in the case of a "cold wave," drop below the 10th percentile. When analyzing features related to consecutive days, it's inherently interesting to determine during which part of the growing season they occur. Since temperature features are defined relative to historical data, they can manifest at any time of the year, not merely during the traditionally hot or cold months. This aspect is further explored in 3.7.

3.6. Limitations

The design of the features described above meets many of our objectives outlined in the first paragraph of this section. However, it is important to acknowledge that the nature of weather data still presents certain challenges in various aspects.

Multicollinearity Minimizing multicollinearity was a significant objective in the process of feature engineering, however it was not always feasible without sacrificing valuable information. In our selected features we anticipate multicollinearity among the features related to the same month, since higher amounts of precipitation lead to lower temperatures. In the adjacent monthly temperature features we are also expecting to observe correlations, as they are likely to display similar weather patterns, particularly during the summer. These summer months are additionally inherently correlated with the "summer days" feature, which offers a broader perspective on the frequency of hot days within a year.

For the monthly precipitation features, we do not foresee multicollinearity as a significant issue, given that the monthly precipitation features quantify the relative deviation from historically recorded precipitation levels. Thus, variations in the general amount of precipitation over the months are not directly included in our features. However, the extreme weather features that count consecutive days are likely to be strongly correlated with the month in which the index is recorded. Nevertheless, since there is no particular indication suggesting for certain month to be more likely to contain the days accounted for by these features, this correlation should not significantly impact our analysis when considering data across all years.

Although we could not completely eliminate multicollinearity in our feature set, we aimed to minimize it as much as possible and are conscious of its presence. We thoroughly examine the correlation structure in 4.1, and exclude features with the highest correlations. Additionally, we will employ techniques specifically designed to address multicollinearity within the statistical analysis to ensure the integrity and validity of our findings.

Trade-off between precision and robust features The balance between creating precise features and selecting variables likely to yield significant effects in statistical analysis, presents an ongoing challenge. From the perspective of precision, particularly in identifying critical periods within the growing process, analyzing smaller time frames, such as weeks instead of months, could offer advantages. This approach allows for the detection of short periods of extreme weather that may significantly impact crops but could be missed due to monthly averaging. For instance, a brief but intense period of heat and dryness within a summer month could severely damage plants, yet this might be overlooked if the rest of the month presents typical weather conditions. Weekly analysis could potentially capture such events more accurately.

However, this approach has several drawbacks. Primarily, since our analysis requires yearly features to predict annual yield, using a single week as a predictor is impractical, as it's unlikely that specific weeks consistently have a disproportionate impact on yield. Furthermore, severe weather events are likely to occur in different weeks each year. Also, adopting a weekly framework would exacerbate the issue at month-end boundaries, where some extreme weather events might already be overlooked if they span over the transition from one month to the next. Additionally, the problem of multicollinearity, previously discussed, would intensify due to the increased number of time intervals.

Therefore, we chose to retain monthly features while introducing additional indices for extreme weather events to capture effects not observed by monthly data. However, these indices still cannot capture every event. For example, we only consider the longest dry spell, which doesn't inform us about the frequency or impact of other, slightly shorter dry spells occurring within the same year. Despite these limitations, we believe our approach

strikes a reasonable balance between reducing multicollinearity, designing meaningful features, and capturing as much relevant information as possible. Nevertheless, we acknowledge that there may be alternative features that better manage these trade-offs and lead to more comprehensive analytical outcomes.

3.7. Descriptive Analysis

The final feature dataframe used in the exploratory analysis as well as the modelling applications consists of 5973 rows and 27 columns, including the target variable *'Canola'* as well as 26 explanatory features derived from the initial information as stored in 2. It is indexed by annual timestamps ranging from 1990 to 2022.

Target Variable The normalized target variable *Canola* centers around a mean -2.997 with a standard deviation of 6.36. In contrast, it has a median of 0.882, indicating left skewness that is confirmed upon visual inspection. This suggests that the majority of the data points are concentrated around the lower end of the distribution, i.e. that a significant portion of the observations have lower-than-average yields.

During the initial analysis, one outlier value was detected in the year 1994 for the rural municipality of Melvin, No. 499. The row was subsequently removed from the dataset.

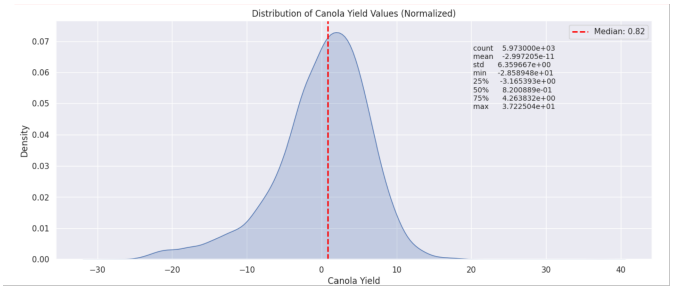


Figure 3. Density plot for the distribution of the target variable *Canola*.

Feature Exploration We begin with an exploratory analysis of the features, where we plot several of the features computed as in 3.5 against one another to obtain an initial visual impression of potential relations in-between them. We are particularly interested in how strongly spatial and temporal differences are observable in the climatic data.

With regard to geographic variations in the temperature, some potentially significant patterns are visible when computing boxplots for each region **??** (note that the x-axis denotes the region ID from 1 to 622 in ascending order). The differences become even more visible when regions are aggregated by cardinal points as illustrated in 4.

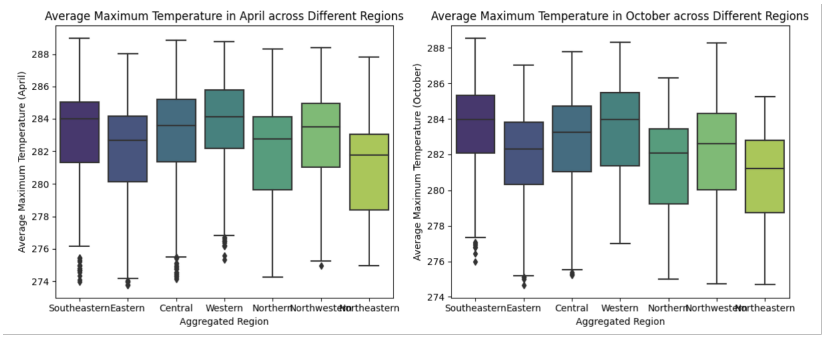


Figure 4. Boxplots for the Average Maximum Temperature in April and October across Grouped Regions.

In contrast, regional fluctuations in the precipitation values appear less significant as 29 in the appendix indicated.

Particularly intriguing insights into the relation between extreme weather events and the yield arose from the line plots in 5. We contrast the occurrence of extremely warm or cold days, either in the context of dry and wet spells or on a stand-alone basis, with annual yield. Days (or spans thereof) with temperatures above 25 $°C$ appear to influence the yield positively up to a certain extent at which the effect reverses. This is in line with prior knowledge of extreme heat stress having a detrimental effect on the plants above a certain threshold. Cold days or spells similarly seem to be tolerated to some extent, but may accelerate the harmful impact of extreme heat in critical years.

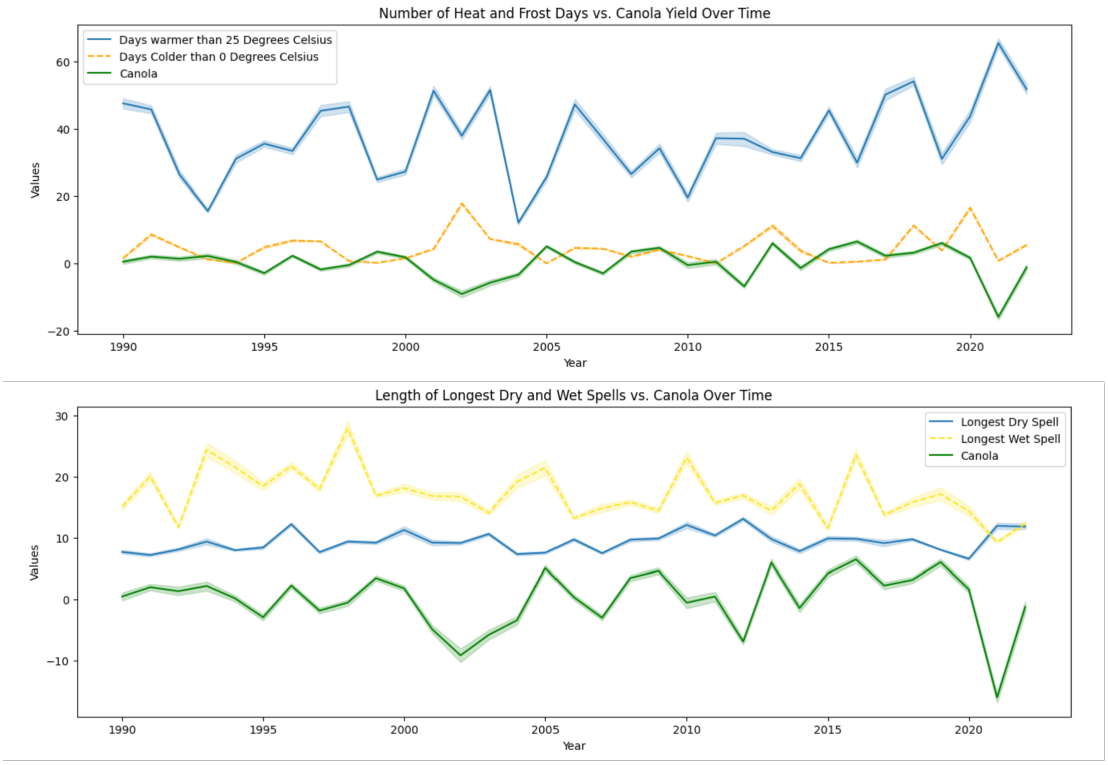


Figure 5. Number of extremely warm and cold days (above) and longest dry and wet spells (below) vs. *Canola*.

Additionally an analysis of the plots in 6 examining the occurrence of extreme weather events over several decades reveals interesting patterns. Notably, over the last 30 years, the longest spells of wet weather have predominantly occurred in June. This finding aligns with the high correlation observed between June's SPI and the longest wet spells 4.1. In contrast, the longest dry spells are most common later in the growing season, a period when drought may actually be favorable for harvesting activities, as excessive moisture can complicate these processes. The longest cold waves have predominantly occurred during spring, with a noteworthy absence in summer, especially since 2010— a pattern consistent with expectations from global warming. Heat waves show a broad distribution across years, surprisingly more common in spring than summer. This may reflect how temperature changes are gauged against historical norms, making a significant warm shift in spring appear extreme compared to smaller increases in already-hot summer months. This perception is also influenced by human tendency to remember absolute high temperatures rather than relative changes.

An issue that had not yet been addressed remained the significance of the soil zone, and whether the mere distinction of black-gray soil from brown and dark brown soil would suffice as a basis to observe and explain potential differences. An exemplary visual plot of Canola in contrast with the average maximum July temperature revealed a clustering tendency to some extent. However, it should be noted that the geographical location of regions in the respective soil zone largely accounts for this as well, as illustrated in 1. There are further limitations in regards

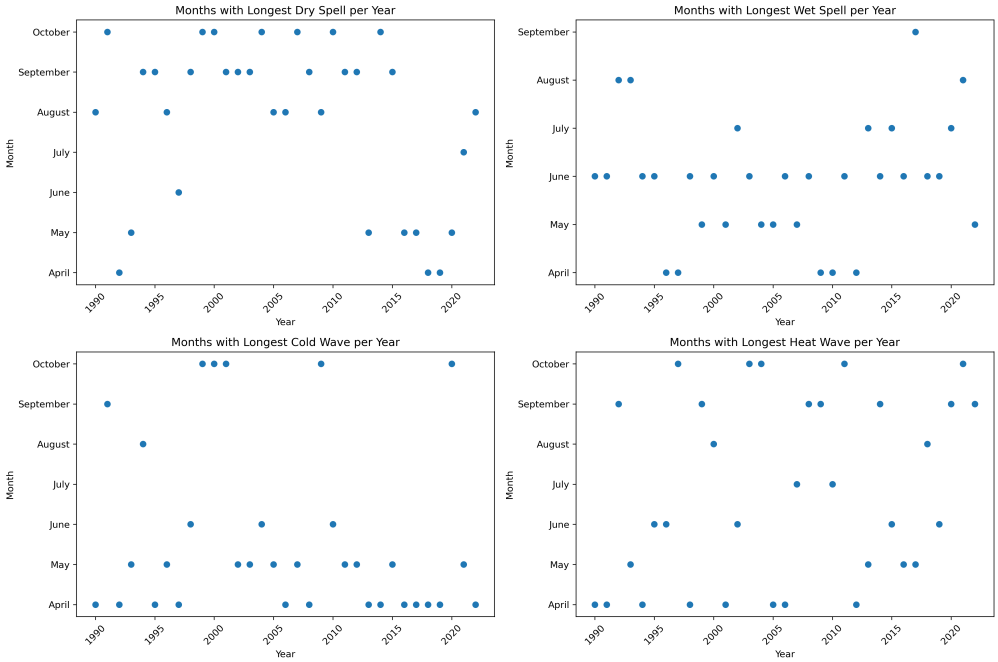


Figure 6. Distribution of the extreme weather events accross the month.

to class balance due to the inconsistent availability of data, as 67.4% of observations stem from zone 1 (black or gray soil) where canola is so far grown primarily, while 28.2% are in zone 2 (dark brown) and only 4.4% in zone 3 (brown) respectively.

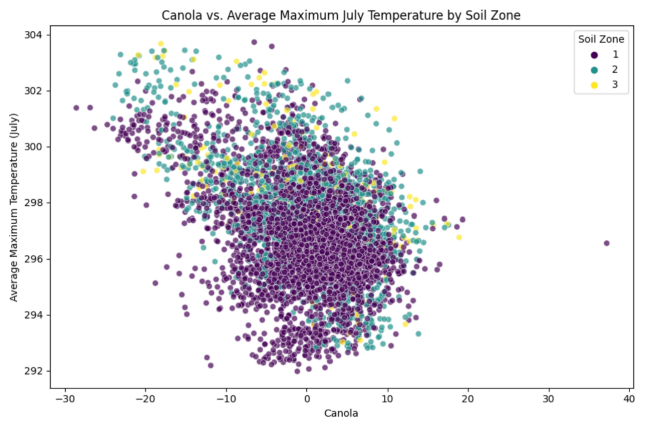


Figure 7. Canola Yield vs. July Temperature by Soil Zone (1 = black/gray, 2 = dark brown, 3 = brown).

4. Results

4.1. Feature Correlations

As previously mentioned in 3.6, complete avoidance of multicollinearity in the construction of features was not feasible. This section delves into the correlation structure among the explanatory features in greater detail. 8 presents a heatmap that illustrates the correlation between the features, including the response variable. The individual Correlation Coefficients are calculated using the Pearson correlation coefficient. The heatmap reveals that many of the concerns regarding feature correlation, as outlined in 3.6, are indeed evident in the actual data.

Weather features For a considerable number of feature combinations, the correlation remains within an acceptable range of -0.25 to 0.25. However, we also observed some instances of very strong correlation, where the absolute value exceeded 0.5. These strong correlations fall into different categories. Firstly, there's a notable correlation between features for the average maximum temperature and the Standardized Precipitation Index (SPI) analyzing the same month. For example, the SPI in September shows a strong negative correlation of -0.66 with the average maximum temperature in September.

Additionally, there is a significant correlation between adjacent months. This is particularly true for the monthly temperature features, while the monthly precipitation features do not seem to encounter this issue. Theoretically, this is plausible since the SPI measures the relative amount of precipitation compared to historical averages for that period, making it less likely to correlate strongly with temperature. An exception to this is the correlation of 0.3 observed for the SPI between May and June. Among the monthly temperature features, the correlation is notably higher, especially during the summer months. The months of June and July display the highest correlation, with a Correlation Coefficient of 0.49. Interestingly, the correlation between May and August is also surprisingly strong, at 0.41, marking one of the strongest correlations among monthly temperature features. In summary, the correlation between temperature and precipitation features within the same month appears to be more problematic than the correlation observed between adjacent months.

We also identified correlations within the extreme weather features, as well as between some extreme weather features and specific months. As discussed in 3.5.2, the methodology used to calculate certain indices inherently predisposes them to correlation with other features. This effect is particularly noticeable for the "summer-days," "frost-days," and "extremely wet days" features, which tally the total number of days meeting specific conditions. It is evident that "summer days" will correlate with the temperatures of summer months, and "frost days" with the coldest months of the growing period, namely April and October. These assumptions are strongly supported by correlation coefficients, revealing a strong negative correlation of -0.57 and -0.6 with the average maximum temperature in April and October, respectively. The "summer-days" feature, counting days with temperatures exceeding 25 degrees, shows an even stronger correlation with temperature features. July and August, being peak summer months, correlate to this feature with coefficients of 0.78 and 0.76, respectively. However, even months not typically classified as summer months, such as May and September, demonstrate significant correlations, with coefficients of 0.46 and 0.52. Overall, the "summer days" feature emerges as the most highly correlated in the entire dataset, underscoring the strong link between specific extreme weather features and the temperatures of corresponding months. The feature counting the extremely wet days shows the lowest levels of correlations from this group but still is remarkably high correlated to the SPI in Mai (0.46), June (0.37) and August (0.37). But also the feature which count the longest consecutive days for some weather event show correlation. The longest heat wave is the most correlated to the features of the month June, for both Temperature (0.42) and Precipitation (0.33), this relates to the examination shown in 6 which states that a majority of the Wet Spells over the last 30 years has happened in June. The longest cold wave seems to relate strongly to the frost days (correlation of 0.51) and in that sense also has a noticeable correlation to the temperature in April and October, just like the feature of the frost days. This also resonated with number of times the longest cold wave fell into these month, especially for April. Although it is surprising that the May feature is not stronger correlated due to the large amount of times the cold wave happened in May. The longest Wet Spell is correlated with SPI of June and the longest Dry Spell seems to be the only feature in the dataset which does not exhibit any correlation to other features of a problematic level.

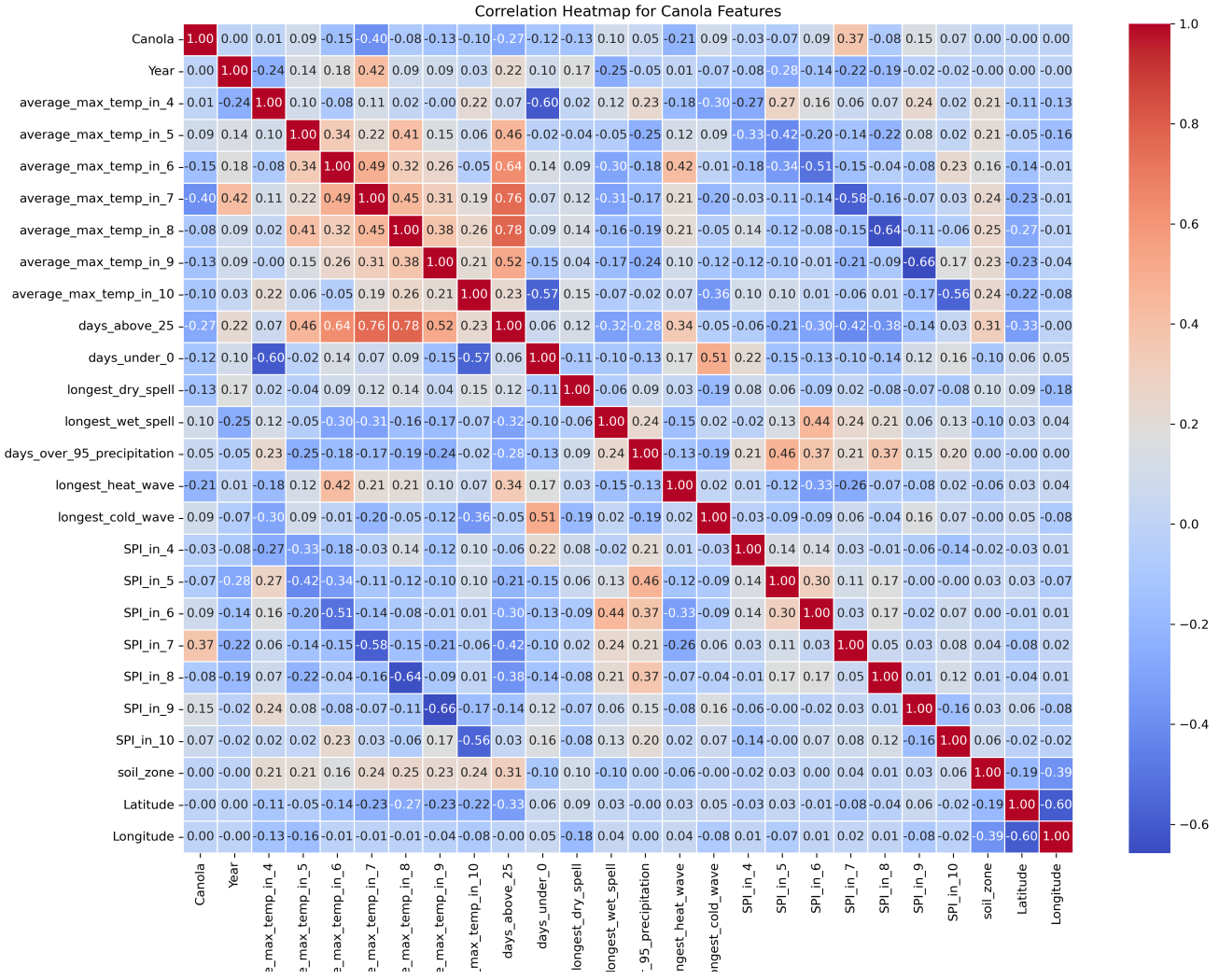


Figure 8. Correlation between features.

Note on Non-Weather Features In our correlation analysis, we also included features that describe the soil zones and the geographical location (longitude and latitude) of the observations. We observed some correlation between the type of soil zones and temperature features. This correlation aligns with the geographical distribution of the soil zones, as illustrated in 7, where, for example, the blue soil zone, located more towards the north, corresponds to a cooler climate. Furthermore, the longitude and latitude values exhibit correlations with the monthly temperature features, suggesting that the climate varies to some extent across the agricultural region of Saskatchewan. This potential spatial influence on climate and, consequently, on agricultural outcomes is further explored in the subsequent section.

Note on response The correlation analysis also provides preliminary insights into which features may be most critical when modeling yield. For the majority of the variables, there is no significant correlation observed, which should not be overly concerning given the basic nature of this correlation analysis. However, the average maximum temperature in July notably stands out with a high correlation coefficient of -0.40. Other features demonstrating some level of correlation include the number of summer days, the duration of the longest heatwave, and the Standardized Precipitation Index (SPI) in July. This suggests that the month of July and high temperatures may play a significant role in influencing the canola yield. These preliminary findings warrant further investigation, as detailed in 4.4.

Exclusion of Variables Due to the high correlations observed, a decision was made to exclude some variables. The most severe correlation issues were identified between the temperature and the Standardized Precipitation Index (SPI) for the same month, and between features that tally weather events for the entire year and those months that typically encompass most of these events.

To mitigate these issues, we decided to remove certain average maximum temperatures that exhibited strong correlations with the SPI indices and the number of summer days, as well as correlations among themselves. Consequently, we excluded the average maximum temperatures for June, August, and October. These months were among the highest correlated features, and we anticipate that their effects should still be indirectly represented through the summer days and SPIs for the corresponding months. Additionally, the feature for extremely wet days was removed, under the assumption that its influence is adequately captured by the SPIs and the longest wet spell feature.

We still observed a very high correlation between the average maximum temperature in July and the number of summer days. However, we deemed both variables too important to exclude. The number of summer days compensates for the removed variables, and the temperature in July is suggested to have the most significant impact on Canola yield.

4.2. Spatial and temporal Dependencies

When analysing data measured over time or space, recognizing inherent dependencies is crucial. Spatial data often exhibit a principle where closer locations tend to show more similarity compared to distant ones. Similarly, temporal data measurements taken in consecutive years are expected to be more alike than those separated by larger time intervals. This section evaluates the extent and significance of these dependencies within our dataset.

The rationale behind the observed correlation in temporal or spatial data necessitates further investigation. In instances of proximity, whether temporal or spatial, it might be that the observed similarities arise due to more closely aligned explanatory features. However, when such alignment does not fully account for the data structure, incorporating temporal or spatial effects becomes crucial. These effects can elucidate variations unexplained by other model variables, and thereby improve our understanding and modelling of the dataset. Given our dataset's structure—comprising measurements taken annually across 30 years at numerous locations—it is plausible to assume inherent temporal and spatial dependencies. This hints at potential autocorrelation structures, including inter-year correlations within a single location, inter-location correlations within a given year, and finally the collective impact of time and space on the dataset's yield across all locations and years.

4.2.1. Autocorrelation Analysis for Time Series In time series analysis, autocorrelation is a key factor. The Autocorrelation Function (ACF) is a standard tool for measuring correlation across different time lags, and determining the significance of autocorrelations within a confidence interval. Our dataset comprises 184 time series corresponding to various regions, alongside an aggregate series for total yield across all regions. Interestingly, temporal autocorrelation appears minimal; only 39 out of the 184 series exhibit significant autocorrelation, that only occurred at the first lag. This minimal temporal influence is further evidenced by the aggregated yield analysis 9, which shows no significant autocorrelation. These observations, particularly the negligible overall effect on cumulative yield, aligns with our assumptions that while prior year weather conditions can influence soil quality and subsequently the yields, immediate weather conditions are likely to have a more substantial impact that overshadows the effect of previous years.

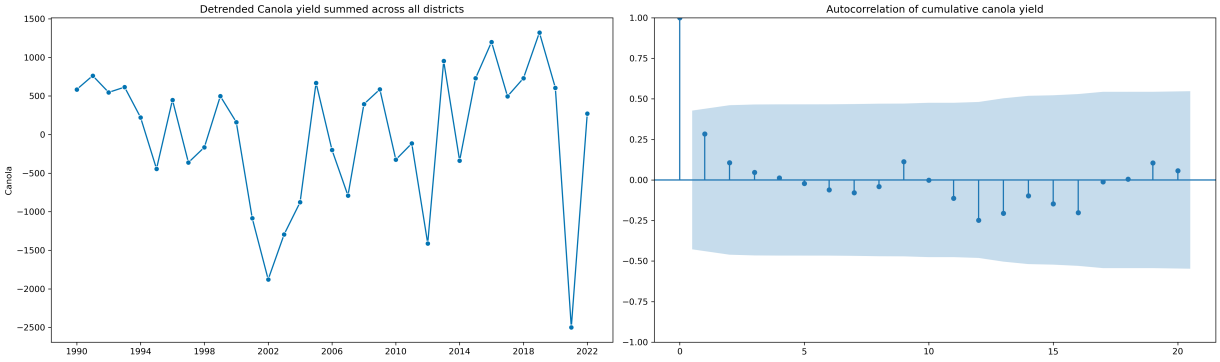


Figure 9. The Time Series (left) and the corresponding Autocorrelation Function (ACF) (right) for the cumulative yield of all districts.

4.2.2. Spatial Autocorrelation and Interpolation Spatial autocorrelation analysis begins with examining yield data modelled against the respective locations. 10 showcases yield interpolation across Saskatchewan's agricultural region using Radial Basis Functions (RBFs) for the overall effect on the cumulative yield. With RBFs the influence of one observation diminish with distance. The values for new location values are calculated from a weighted sum of the known observations. The interpolation was conducted annually and for the cumulative yield using thin plate splines as radial basis function. Here it is crucial to use the original time series with its inherent trend for the cumulative yield analysis to derive meaningful outcomes. Notably, spatial autocorrelation is evident, as shown by the pronounced spatial differences in the interpolation of the cumulative yields in 10. This observation suggests robust spatial dependencies persisting across years. Similar patterns of significant variation are observable in the data for individual years, yet the precise arrangement of the locations with stronger and weaker yield clusters varies considerably.

Further examination of spatial dependence is afforded by variograms, which assess the variance between points relative to their separation distance. 10 also presents the variogram for cumulative yield, where black points symbolize the empirical variogram calculated at specified lags. The theoretical variogram, represented by the red line, is extrapolated from the empirical data, incorporating assumptions about the underlying Gaussian random field, such as its correlation function, with an exponential model employed for this analysis. The variogram reveals a steep increase in variance at shorter distances, which stabilizes as the distance extends, albeit with some fluctuations at the farthest lags. However, the reliability of these distant lag observations is compromised by their smaller sample sizes. While variograms for individual years vary, a general trend of increasing variance with distance emerges, supporting the spatial dependence insights obtained from interpolation.

The question that now emerges is whether the observed autocorrelation can be accounted for by the existing features, or if there exists an additional spatial effect that could elucidate the variance not explained by the explanatory variables. Addressing the potential for explanatory variables to account for observed autocorrelation, Moran's I was applied to each year's data. As outlined in Section 2.2, Moran's I can be used to quantify autocorrelation. Here it is used in two distinct ways: first, to assess the overall autocorrelation, and second, to evaluate the residual autocorrelation after accounting for the effects of the explanatory variables. **??** displays these findings, with the general yearly autocorrelation (blue line) indicating notable clustering effects in many years, and a mean autocorrelation of 0.43. However, certain years exhibit exceptionally high autocorrelation, surpassing 0.7. The red line represents the portion of autocorrelation that remains unexplained by the explanatory variables and the application of OLS Regression. Predominantly, this value fluctuates between 0.0 and 0.1, indicating it remains only a minor fraction of the initial autocorrelation. This suggests the majority of autocorrelation can be attributed to the variables considered, without the introduction of an additional spatial effect. Utilizing Moran's I, it is also feasible to obtain p-values, that indicate whether the residual autocorrelation is due to an extraneous spatial factor. Significance in p-values was observed in only 9 of the 30 years examined. Even in cases with significant p-values, the magnitude of unexplained autocorrelation remains sufficiently minor to warrant no further inquiry. An

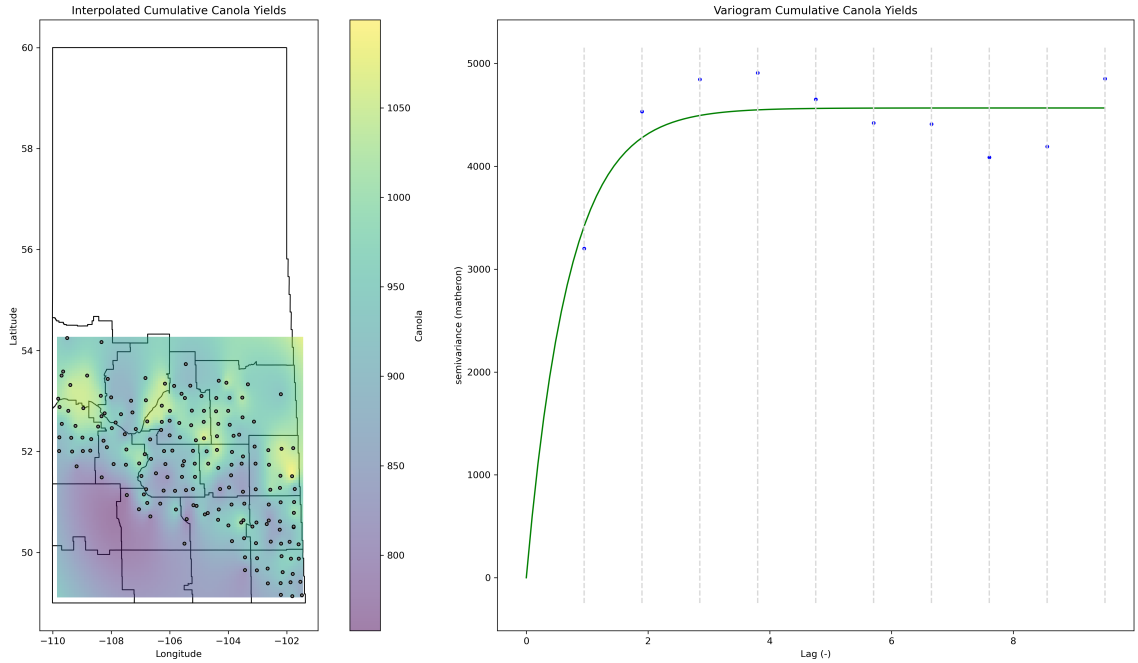


Figure 10. The yield interpolated over the agricultural region of Saskatchewan (left) and the corresponding Variogram (right) for the cumulative yield over all years.

exception is noted for the year 2022, where an unexplained autocorrelation exceeding 0.2 was recorded, coinciding with a notably high overall autocorrelation for that year.

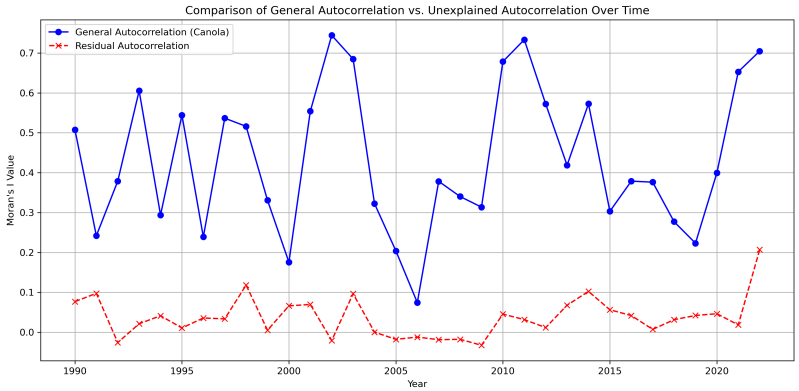


Figure 11. The total autocorrelation compared to the unexplained autocorrelation based on Moran's I

4.2.3. Conclusion Our comprehensive analysis of temporal and spatial dependencies reveals minimal temporal autocorrelation when evaluating total yield across all regions. Conversely, substantial spatial autocorrelation is identified for individual years and the cumulative effect. However, the majority of this autocorrelation is attributable to local weather variations. Consequently, while autocorrelation in our dataset is not negligible, it should not dominate our analytical focus, allowing for a balanced approach for model developments.

4.3. *Principal Component Analysis*

Computation of principle components was initialized using a randomized approach to Singular Value Decomposition (SVD), intending to optimize computational efficiency. This yielded principal components surpassing 80% of cumulative explained variance at the eighth component as illustrated in 12. Aiming for an explained variance between 80% and 85%, we opted to move forward taking 10 principle components into account (81.4%).

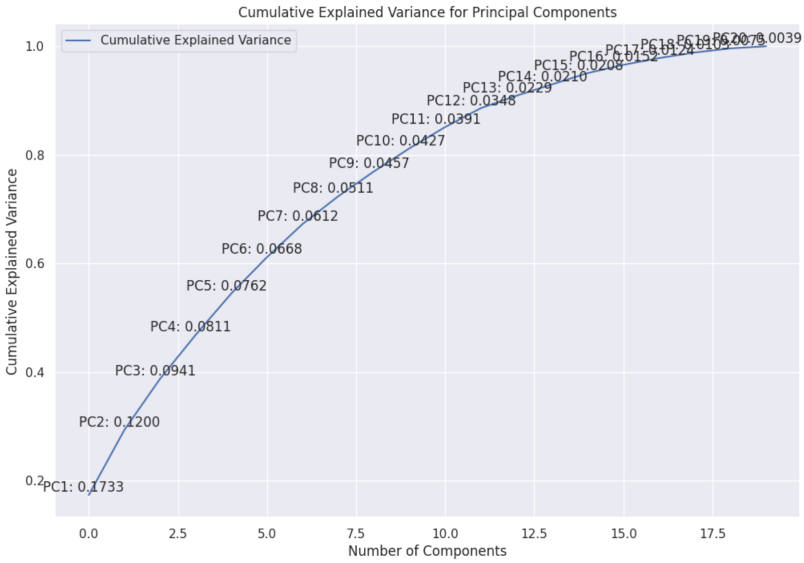


Figure 12. Explained Variance Portions by Principle Component.

Feature mappings among the four most significant principle components, i.e. those explaining the largest respective portions of variance, provide an insight into how variable relations were handled and can be found in A.

Visualizing the correlation heatmap in 13 between the components confirmed the successful removal of all critical correlations between features.

To see how the original variables are distributed in the components, a heatmap displaying the absolute magnitude of all feature coefficients within each component was computed 14. Notably, PC1 predominantly captures variability related to seasonal temperature patterns. The dominant variable by coefficient is *days_above_25*, indicating the large influence of extreme heat stress on fluctuations in Canola yield. An increase in PC1 is furthermore associated with higher temperatures in May, July, and September, as well as shorter durations of wet spells and longer durations of dry and heat spells.

The second principal component primarily captures temperature characteristics in the earlier stage of the canola life cycle, with larger absolute coefficients for *average_max_temp_in_4* along with *days_under_0* and *longest_cold_wave*. This excellently illustrates the function of individual components to capture correlated variables and subsequently account for the removal of multicollinearity. This component also considers the variable *soil_zone* more strongly than the first one does. This is in line with the domain understanding that the characteristics of the soil have a particularly large impact on Canola in its early growth stages, especially during seedling emergence and establishment [16]. Other factors contribute rather moderately.

PC3 primarily captures geographical variables and the inevitable correlations therein as well as the average maximum temperature as well as the standard precipitation index in the month of September. *average_max_temp_in_4* as well as *average_max_temp_in_5* are also weighted considerably. Other notable coefficient magnitudes include *SPI_in_7* in the seventh principle component (31.95%, explained variance of component: 6.12%); *SPI_in_6*, *SPI_in_8* and *longest_heat_wave* in the eighth principal component

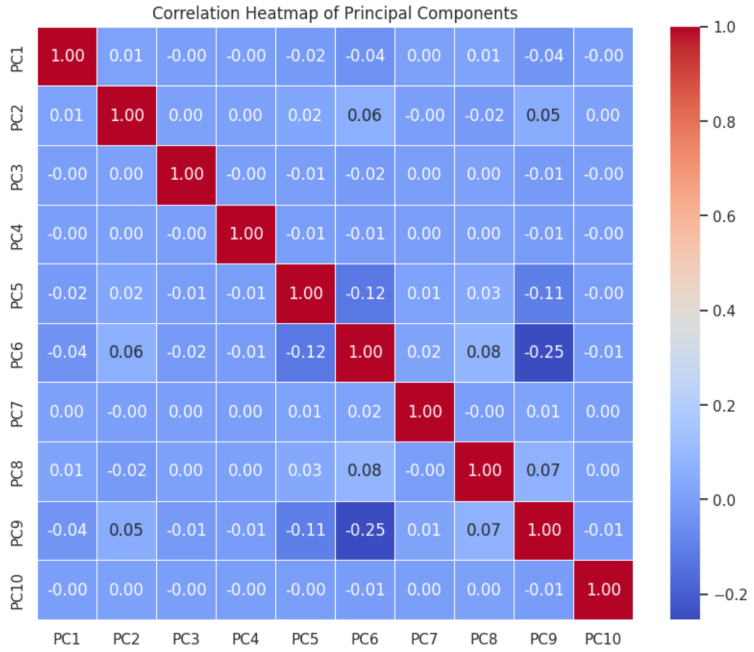


Figure 13. Heatmap of Correlations between Principle Components.

(21.73%, 23.82%, and 12.57% respectively, explained variance of component: 5.11%); *longest_dry_spell* and *longest_wet_spell* in the ninth principal component; and *SPI_in_5* and *SPI_in_8* in the tenth principal component.

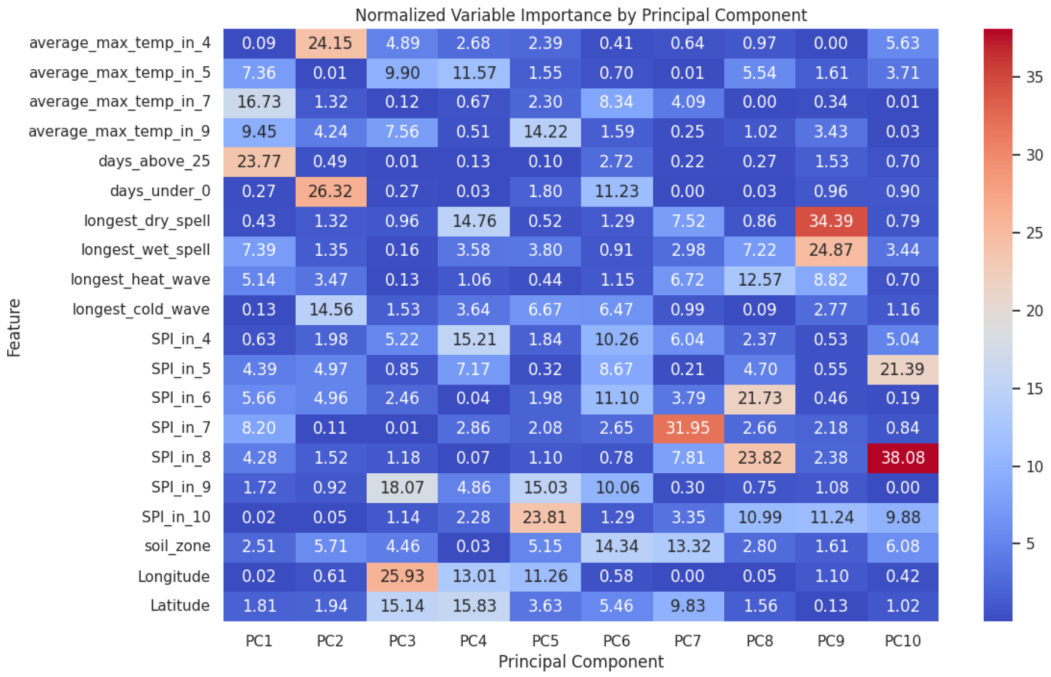


Figure 14. Feature Magnitude Heatmap in Principle Components.

4.4. Modelling Results

Modelling applications were split into two parts based on the explanatory features going into the models. In the first section, we use the variables constructed as detailed in 3.5 and selected as in 4.1. For models in the second section, we use the ten first principle components in accordance with 4.3. In each approach, we fit a selection of traditional regression models as well as three machine learning approaches to investigate their fitness for the data.

4.4.1. Initial Variables Prior to fitting the models, we split our data into a training set and a test set with a 80% to 20% ratio. We further scale the feature sets to ensure consistency.

Linear and Ridge Regression Linear and Ridge Regression are implemented first, with *alpha* set to 0.1 for the latter. Only deviating marginally from one another, both linear and Ridge regression achieve R^2 values of 31.46% with mean squared errors of 27.15.

Model	Parameter	R^2	MSE
Linear Regression		31.46%	27.15
Ridge Regression	$\alpha = 0.1$	31.46%	27.15

Table 4. Comparison of Linear and Ridge Regression

LASSO Regression LASSO regression is instantiated using a cross-validated *alpha* of 0.01. It yields slightly worse results with an R^2 of 31% and a mean squared error of 27.16, but is worthy of being looked into further for information on variable importance and the selection process.

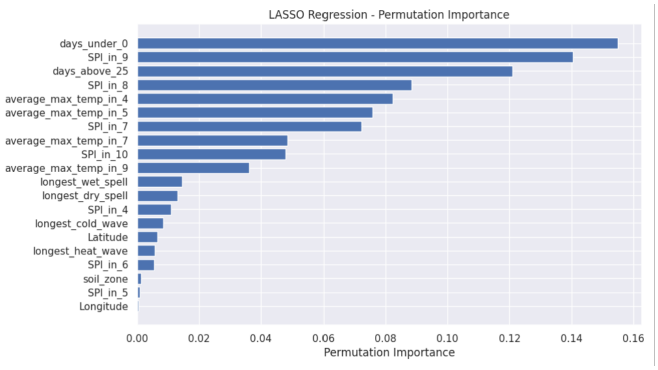


Figure 15. Permutation Importance for Variable-Based LASSO Regression.

Features of notable magnitude include the extreme temperature events as well as the average maximum temperatures in May and July. The standard precipitation index is of notable relevance in the months from July through September, suggesting a particular need for adequate precipitation during the flowering period. Little importance is attributed to geographical variables.

Support Vector Regression The initial SVM is fitted using default settings, i.e. an RBF kernel along with a scaled *gamma* and a regularizing *C* of 1. This results in a MSE of 12.89 and a R^2 of 67%. While these numbers demonstrate significantly improved predictive performance, the visualization of the predictions in 17 reveals a strong non-linear pattern in the residuals. This points to either a misspecification of the model or the absence of vital information in the input variables. While the latter supposition is hard to evaluate, we attempt to account for potential misspecification through a parameter optimization approach in 4.4.2.

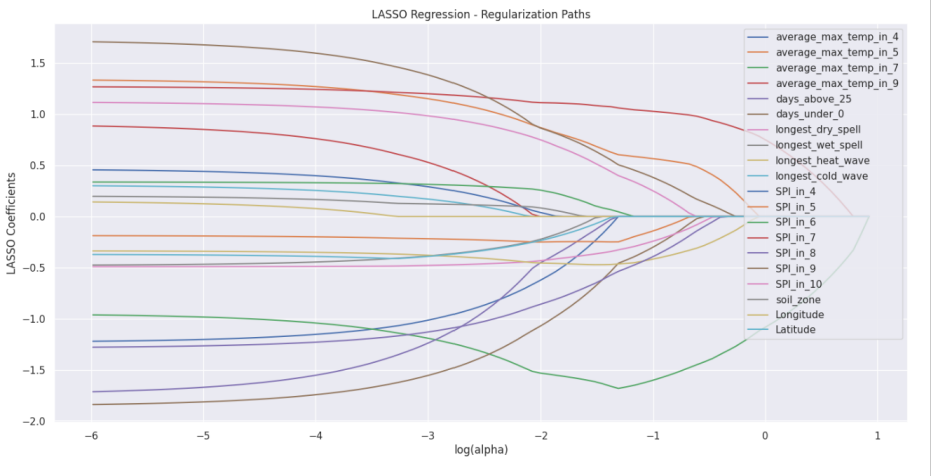


Figure 16. Regularization Paths for Variable-Based LASSO Regression.

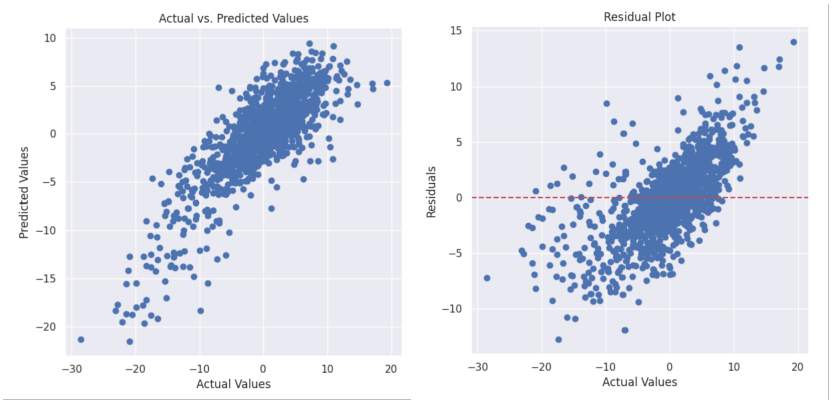


Figure 17. Predictions and Residuals for Variable-Based SVR.

Random Forest Regression A Random Forest is instantiated using default settings.

1. *n_estimators*: 100
2. *criterion*: squared_error
3. *max_depth*: None
4. *min_samples_split*: 2
5. *min_samples_leaf*: 1
6. *max_samples*: None

Numerically, this yields slightly better results than the previous approach, namely an MSE of 10.16 and an R^2 of 74%. Visualizations reveal similar, albeit slightly weaker tendencies as before, particularly with regard to the non-linear residual pattern. Again, we explore the model's potential to account for this using an optimization technique detailed in 4.4.2.

In the context of variable importance, the random forest places more importance on the average maximum temperature than the previous LASSO regression, which is in line with what we expected as the temperature around this early flowering stage is known to be crucial. The second and third most important features are the precipitation in May and the number of extremely cold days. Again, climate-related variables overall display significantly higher scores than those of geographical nature.

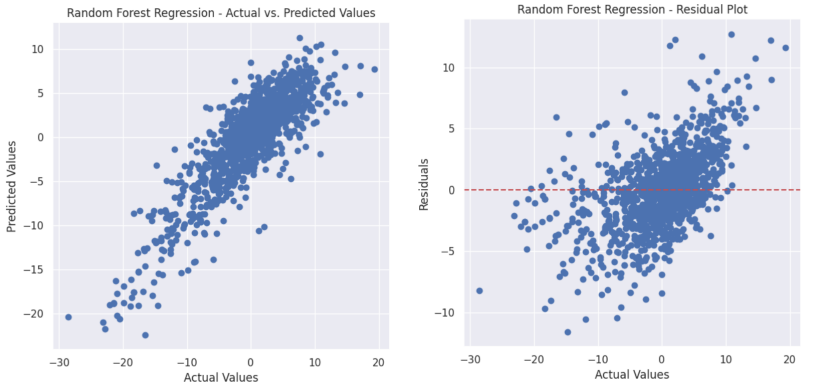


Figure 18. Predictions and Residuals for Variable-Based Random Forest.

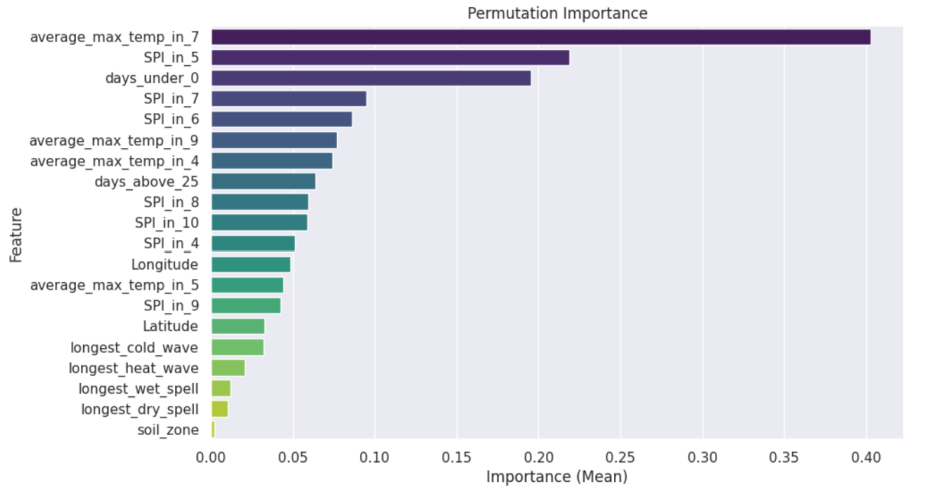


Figure 19. Permutation Importance Scores for Variable-Based Random Forest Regression

LSTM Finally, the neural network architecture is implemented using LSTM. It consists of a LSTM layer activated through the tanh function as well as a Dense output layer, and is compiled using the Adam optimizer and MSE as the loss criterion. Having trained for 80 epochs, it comes in at an MSE of 10.68 and a R^2 of 73%.

4.4.2. Principle Components Hoping to increase model accuracy and resolve the issue caused by non-linearity in the residuals, we train the same models as previously using the derived principle components as input features. The data is again scaled and split into 80% training and 20% test sets respectively. We further implement cross validation to optimize some model parameters.

Linear and Ridge Regression Linear Regression and Ridge Regression again perform similarly. Both models produce a mean squared error of 30.86 while achieving R^2 values of 22.09%, showcasing slight degradations in performance as opposed to the models trained on non-reduced input features. The choice of *alpha* in the Ridge regression model did not affect its predictive performance.

LASSO Regression We implement the first variation of LASSO regression using 0.1 for *alpha*, yielding analogous results to Linear and Ridge Regression. Upon cross validation, it was deemed optimal at 0.035 and thus

Figure 20. Predictions and Training Loss History for Variable-Based LSTM.

indicated little need for regularization. A plot illustrating the MSE development while determining the optimal value for *alpha* can be found in the appendix A.

Notably, while the seventh principle component accounts for only 6.12% of explained variance, it is considered heavily by the LASSO regression model (-1.34). The dominant variable within this component is *SPI_in_7*, signaling a high importance of July precipitation in the models' decision. The second and third most influential components are PC5 and PC8 with coefficients of -0.95 and 0.8 respectively.

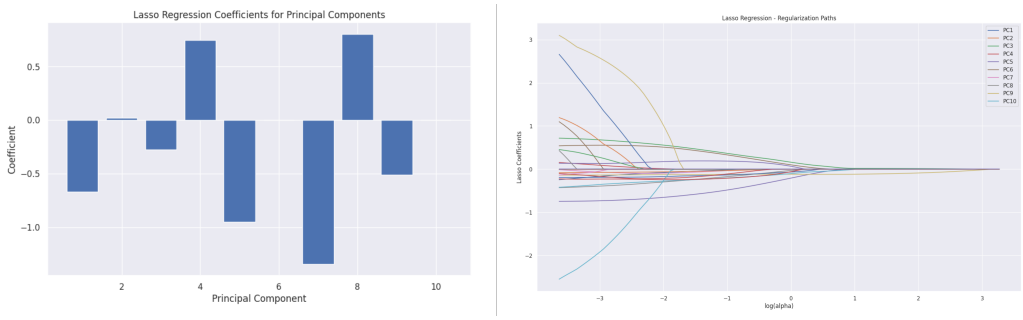


Figure 21. Coefficient Magnitude and Regularization by Component in PC-based LASSO Regression.

Support Vector Regression Similarly to LASSO, we begin with implementing a SVR model using a RBF kernel with no further parameter specifications and rely on the default settings 'scale' for *gamma* and 1 for *C*. It yields a R^2 of 63.98% with an MSE of 14.26, a degradation from the initial variable-based model detailed in 4.4.1 of around 3% in explained variability and an increase of 1.37 in the error. Visual inspection of the model predictions 22 reveals the persistence of the non-linear pattern in the residual values.

In an attempt to relieve this issue and improve predictive performance, we implement a grid-based cross validation approach to optimize *C* as well as *gamma*. We prompt the model to test all combinations of *gamma* being 0.001, 0.01, 0.1, 1 or 10, and *C* of 0.1, 1, 10, 100, or 1000 respectively. Optimal parameters are determined at $gamma = 0.1$ and $C = 10$.

A model implementation using the optimized parameters yields an R^2 of 72.35% and a MSE of 10.94%. Visual inspection of the updated predictions shows a significant improvement in the removal of the pattern from the residuals 23.

Figure 22. Actual vs. Predicted and Residual Values for default PC-based Support Vector Regression.

Figure 23. Actual vs. Predicted and Residual Values for cross validated PC-based Support Vector Regression.

Random Forest Regression Building on the success of the previous cross-validation approach, we implement it analogously for a Random Forest regression. Optimized parameters return as $'max\_depth'$: $None,'max\_samples'$: $0.9,'min\_samples\_leaf'$: $2,'n\_estimators'$: 300. We thereby achieve a similar score of 71.08% in explained variability and a MSE of 11.45.

Figure 24. Actual vs. Predicted and Residual Values for cross-validated PC-based Random Forest Regression.

LSTM A LSTM network is defined as a sequence of layers, consisting of a single LSTM layer with 50 units as well as a dense output layer for the prediction. The Adam optimizer is applied and MSE prompted as the loss function. We opt for a batch size of 32.The intial run is conducted using 200 epochs to gain a sufficiently broad insight into the number of epochs actually needed.

The lowest test loss occurred at epoch 102, achieving a MSE of 13.55 and a R^2 of 63.53%. This prompts us to re-train the model using a maximum number of 130 epochs and with an additional early stopping mechanism with a patience of 10 implemented.

With early stopping at epoch 116, a test loss of 13.52 is achieved in this second run.

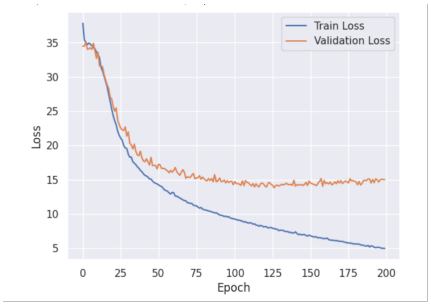


Figure 25. History of LSTM training and test losses over 200 epochs.

Figure 26. Actual vs. Predicted and Residual Values for PC-based LSTM Network.

4.4.3. Remarks Dimensionality reduction of the initial variables did not lead to modelling improvements in direct comparison, with even some minor concessions made in predictive performance. However, the approach still proved to be useful, as it allowed for in-depth analyses of intra-variable relations and provided a perspective on model decision making contrasted between original and transformed variables. Moreover, using principle components in combination with optimzation techniques for model parameters turned out to be the approach with which non-linear residual patterns were best accounted for.

Although a principal component reduction of the initial partially correlated variables yielded rather small portions of explained variance for components beyond the sixth, it proved to be sensible to include components up to at least the eighth in various modelling applications, as the latter ones played significant roles in some of the models' decision-making processes. Particularly the seventh principle components stands out in both the LASSO and Random Forest implementation.

The cross-validated parameter values in Support Vector Regression imply a certain degree of noise in the data and reduce the sensitivity to individual data points such that the model generalizes better. The moderately high value for C accounts for the balance between bias and variance. While the non-linear pattern was largely removed, further deviations of the predictions from the actual values were still observable around the end of the tails, implying persisting issues with the prediction of extreme values. This is likely partially attributable to the smaller sample size in these ranges relative to the middle part.

While overfitting was rather well-accounted for in the SVR implementation, the cross-validation approach steered the Random Forest towards this behavior. The number of trees in the model was always determined at the maximum possible value, while maximum depth of individual trees was never limited and the minimum number of samples required at a leaf split was chosen as small as possible. Visualization of the predictions demonstrated overly narrow patterns within the training data in contrast to the test set. In the context of permutation importance displayed in 27, the random forest relied largely on the first principle component, which is in accordance with its large portion of explained variance in the data (17.55%). Notably similar importance is attributed to the fifth principle component, which emphasized climatic conditions in September and October. Again, the seventh component stood out with the third largest importance.

The LSTM network was computed with highest possible simplicity in its architecture. This led to competitive predictive results and has the advantage of high transparency. While an extension to a more complex architecture

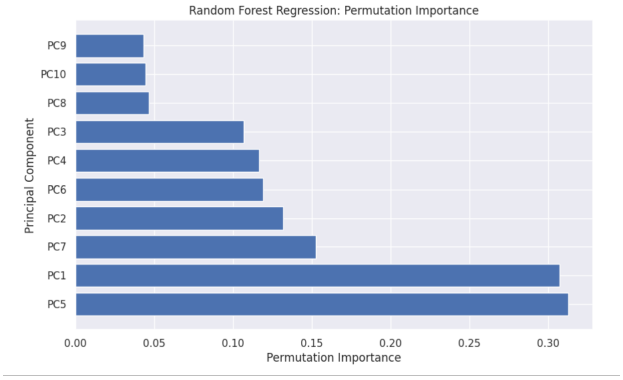


Figure 27. Permutation Importance in cross-validated RF Regression.

(e.g., through adding additional layers) may appear intuitive in search of higher accuracy, several attempts to do so as well as to adjust the parameters detailed in 2.2 actually resulted in worse predictive performance. We subsequently settled on the simple architecture and initial parameters chosen.

5. Discussion

While optimized machine learning techniques show promising results in handling multicollinearity, further reducing multicollinearity among our features could open up possibilities for exploring additional statistical methods. Mixed models, ideal for our data's clustered structure with crossed random effects for years and regions, were considered. However, these methods did not yield promising results, leading us not to pursue them further. The strong multicollinearity among the fixed effects likely poses challenges for these statistical approaches. Effectively managing multicollinearity in the dataset could broaden the range of statistical methods applicable for analysis. By reducing multicollinearity, it becomes possible to leverage a wider variety of analytical techniques, potentially leading to additional insightful and robust findings.

Addressing multicollinearity could benefit from innovative feature engineering strategies. Notably, the monthly SPI features exhibit less correlation compared to temperature features. Modeling temperature features to reflect deviations from historically observed temperatures might offer a similar advantage. This approach could enhance model interpretability and effectiveness by focusing on relative changes rather than absolute values, potentially mitigating issues related to multicollinearity.

Further innovative feature engineering strategies could help us to addressing the multicollinearity. Notably, the monthly SPI features exhibit less correlation compared to temperature features. Modeling temperature features to reflect deviations from historically observed temperatures might offer a similar advantage. This approach could enhance model interpretability and effectiveness by focusing on relative changes rather than absolute values, potentially mitigating issues related to multicollinearity. Although defining such features for the temperature might be less straightforward as for the precipitation.

Given our findings that climatic conditions in July significantly affect canola growth, further exploration into the timing of weather events could be of value. One method could involve analyzing features related to the duration of weather events and their occurrence within specific months. As already discussed in 3.6 our current metrics may overlook brief, intense events crucial for impact assessment. Developing features that capture short, extreme weather events, such as consecutive days with high temperatures, and considering their interaction with other factors like preceding drought conditions, could offer deeper insights into their agricultural implications. One approach here could be to construct features which take larger time spans into account and count short but extreme events, e.g. three days in a row exceeding temperatures of 30 degrees. Maybe with these kind of features even the interaction effect between heat waves and drought in a certain time spans could be investigated.

6. Conclusion

In our modeling attempts, we identified that Support Vector Regression, optimized through cross-validation and paired with scaled principal components, outperforms other techniques. It demonstrates both the highest predictive accuracy and robust generalization capabilities. Consequently, we advocate for the adoption of this model alongside the computation of average maximum temperatures and standard precipitation indices during the key months of the crop's life cycle. These variables are then mapped to principal components, establishing a standardized data processing pipeline for future yield modeling efforts. This approach is particularly beneficial for crops sensitive to temperature, precipitation, and soil conditions.

In our study, using optimized machine learning techniques exhibit promising results, especially in managing the multicollinearity among our features. Nevertheless our study underscores the importance of addressing multicollinearity in agricultural yield modeling to unlock the full potential of advanced statistical methods. Through further reduction of collinear features the scope for exploring alternative statistical approaches could be expanded. Moving forward, further research into constructing features that encompass larger time spans, potentially revealing interaction effects between heatwaves and droughts, holds promise for advancing our understanding of crop-yield dynamics.

Through this work, we particularly hope to empower farmers and other agricultural stakeholders through valuable insights into potential harvest outcomes, enabling better-informed decision-making regarding crop management practices and resource allocation.

REFERENCES

1. Abdi, H., Williams, L. J., *Principal component analysis*. WIREs Computational Statistics, 2(4), 433-459, 2010.
2. Agriculture and Agri-Food Canada, *Government invests in precision agriculture to enhance competitiveness and efficiency*, Canada.ca. https://www.canada.ca/en/agriculture-agri-food/news/2022/02/government-invests-in-precision-agriculture-to-enhance-competiveness-and-efficiency.html, 2022.
3. *Precipitation*, Glossary of Meteorology, American Meteorological Society, 2009.
4. K.Bahadur KC, D. Montocchio, A. Berg et al., *How Climatic and Sociotechnical Factors Influence Crop Production: A Case Study of Canola Production*, *SN Appl. Sci.*, vol. 2, p. 2063, 2020.
5. BASF, *Benefits for Canola Crops*, BASF Crop Protection. Accessed on March 15, 2024. URL: https://agriculture.basf.com/global/en/business-areas/crop-protection-and-seeds/weed-management/glufosinate-ammonium/benefits/benefits-for-the-crops/canola.html.
6. A. Beck, and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.
7. W. J. van de Berg, M. R. van den Broeke, E. van Meijgaard, F. Kaspar, *Importance of Precipitation Seasonality for the Interpretation of Eemian Ice Core Isotope Records from Greenland*, *Clim. Past*, vol. 9, pp. 1589–1600, 2013.
8. Government of Canada, Statistics Canada. *Saskatchewan continues to live up to the title of breadbasket of Canada*. https://www150.statcan.gc.ca/n1/pub/96-325-x/2021001/article/00008-eng.htm, 2021.
9. L. Breiman, *Random Forests*, *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001.
10. Canola Council of Canada, *Precision agriculture*, Canola Council of Canada. https://www.canolacouncil.org/canola-encyclopedia/precision-agriculture/. 2022.
11. E. Candés, and Y. Plan, *Near-ideal model selection by L1 minimization*, Annals of Statistics, vol. 37, pp. 2145–2177, 2008.
12. E. Candés, and J. Romberg, *Practical signal recovery from random projections*, Wavelet Applications in Signal and Image Processing XI, Proc. SPIE Conf. 5914, 2004.
13. SaskCanola. *Canola Industry*. SaskCanola. https://www.saskcanola.com/canola-industry, accessed on March 15, 2024.
14. Institut climatique du Canada. *Heat waves in Canada - Extreme heat in Canada*. Canadian Climate Institute. https://climateinstitute.ca/reports/extreme-heat-in-canada/, 2024.
15. E. Candés, J. Romberg, and T. Tao, *Robust uncertainty rinciples: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 489–509, 2006.
16. Canola Council of Canada. *Effects of Soil Characteristics — Canola Encyclopedia*. https://www.canolacouncil.org/canola-encyclopedia/field-characteristics/effects-of-soil-characteristics/, 2022
17. A. Chambolle, and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numerische Mathematik, vol. 76, pp. 167–188, 1997.
18. T. F. Chan, and S. Esedoglu, *Aspects of total variation regularized ℓ_1 function approximation*, SIAM Journal on Applied Mathematics, vol. 65, pp. 1817–1837, 2005.
19. T. F. Chan, S. Esedoglu, F. Park, and A. Yip, *Total variation image restoration: Overview and recent developments*, in Handbook of Mathematical Models in Computer Vision, edited by N. Paragios, Y. Chen, and O. Faugeras, Springer-Verlag, New York, pp. 17–31,

2006.
20. J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, W.-K. Cheng, S. W. Phoong, Z.-W. Hong, Y.-L. Chen, *Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review*, *Mathematics*, 2022.
21. M. Clancy, *Is technological progress in US agriculture slowing?*, Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris, 2023.
22. Copernicus Climate Change Service *cds.climate.copernicus.eu* https://cds.climate.copernicus.eu, 2023
23. Canola Council of Canada. *Canola production statistics*. Canola Council of Canada. https://www.canolacouncil.org/markets-stats/production/, 2024.
24. Canola Council of Canada. *Canola industry in Canada, from farm to global markets*. https://www.canolacouncil.org/about-canola/industry/, 2023.
25. D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.
26. EOS Data Analytics. *Canola Growing: Planting, Care and Harvesting for High Yield*. https://eos.com/blog/growing-canola/, 2023.
27. Fahrmeir, Ludwig and Kneib, Thomas and Lang, Stefan and Marx, Brian D. *Regression: Models, Methods and Applications* Springer, 2015.
28. *World Food and Agriculture – Statistical Yearbook 2022*, FAO eBooks. https://doi.org/10.4060/cc2211en, 2022.
29. FAO. *Agricultural production statistics. 2000–2021.* FAOSTAT Analytical Brief Series, No. 60, Rome. https://doi.org/10.4060/cc3751en, 2022.
30. S. Hochreiter, J. Schmidhuber, *LSTM Can Solve Hard Long Time Lag Problems*, *Advances in Neural Information Processing Systems*, pp. 473–479, Jan. 01, 1996.
31. Canada, Employment and Social Development. *Saskatchewan Sector Profile: Agriculture - Job Bank*. https://www.jobbank.gc.ca/trend-analysis/job-market-reports/saskatchewan/sectoral-profile-agriculture, 2023.
32. I. Jolliffe, *Principal Component Analysis*, In: M. Lovric (eds), *International Encyclopedia of Statistical Science*, Springer, Berlin, Heidelberg, 2011.
33. Karl, T. R., Nicholls, N., and Ghazi, A., *Clivar/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary*, in Weather and Climate Extremes (Dordrecht: Springer) 3p. –7., 1999
34. Klompenburg, T. van, Kassahun, A., Catal, C., *Crop yield prediction using machine learning: A systematic literature review*, Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.
35. Konduri VS, Vandal TJ, Ganguly S and Ganguly AR, *Data Science for Weather Impacts on Crop Yield*, Frontiers in Sustainable Food Systems, p. 4,52, 2020 https://www.frontiersin.org/articles/10.3389/fsufs.2020.00052/full, accessed on March 15, 2024.
36. M. Liu, S. Hu, Y. Ge, G. B. M. Heuvelink, Z. Ren, X. Huang, *Using Multiple Linear Regression and Random Forests to Identify Spatial Poverty Determinants in Rural China*, *Spatial Statistics*, vol. 42, Article 100461, 2021.
37. Q. Lu, S. Sun, H. Duan, S. Wang, *Analysis and Forecasting of Crude Oil Price based on the Variable Selection-LSTM Integrated Model*, *Energy Informatics*, vol. 4, no. 2, pp. 47, Sep. 24, 2021.
38. Monteiro, A., Santos, S., Gonçalves, P., *Precision Agriculture for Crop and Livestock Farming-Brief Review*, Animals (Basel), vol. 11, no. 8, p. 2345, 2021.
39. Moraga, Paula *Spatial Statistics for Data Science: Theory and Practice with R*. Chapman & Hall/CRC Data Science Series, 2023.
40. H. Y. Osrof, T. C. Ling, A. Gunasekaran, S. F. Yeo, K. H. Tan, *Adoption of smart farming technologies in field operations: A systematic review and future research agenda*, *Technology in Society*, vol. 75, p. 102400, 2023.
41. N. Parveen, S. Zaidi, M. Danish, *Support Vector Regression Model for Predicting the Sorption Capacity of Lead (II)*, *Perspectives in Science*, *Recent Trends in Engineering and Material Sciences*, vol. 8, pp. 629-631, 2016.
42. The Government of Saskatchewan *dashboard.saskatchewan.ca* https://dashboard.saskatchewan.ca/agriculture/rm-yields/rm-yields-data, 2024.
43. SaskCanola. *About Canola*. SaskCanola. https://www.saskcanola.com/about-canola, accessed on March 15, 2024.
44. *Saskatchewan soil zones*. SCIC. https://www.scic.ca/resources/maps/saskatchewan-soil-zones, accessed on March 15, 2024.
45. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
46. Song, Z., Cheng, F., Zhang, Y., *Study on Precision Agriculture Knowledge Presentation with Ontology*, AASRI Procedia, vol. 3, pp. 732-738, 2012.
47. Copernicus European Drought Observatory (EDO) *EDO INDICATOR FACTSHEET - Standardized Precipitation Index (SPI)* https://edo.jrc.ec.europa.eu/documents/factsheets/factsheet_spi.pdf, 2020
48. Statista. Global oilseed production 2023/24, by type. https://www.statista.com/statistics/267271/worldwide-oilseed-production-since-2008/, 2024.
49. Staudemeyer, R. C., Rothstein, E., *Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks*. CoRR. vol. abs/1909.09586, 2019.
50. Government of Canada, Statistics Canada. *November estimates of production of principal field crops*. https://www150.statcan.gc.ca/n1/daily-quotidien/221202/t001b-eng.htm, 2022.
51. Timlick, J., *More canola in the Brown soil zone?*, Grainews.ca. https://www.grainews.ca/features/more-canola-in-the-brown-soil-zone/, 2023.
52. United Nations. *Population — United Nations*. https://www.un.org/en/global-issues/population.
53. V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn., pp. 138–147. Springer, Berlin, 1999.
54. Webster, Richard, and Margaret A. Oliver., *Geostatistics for environmental scientists*, John Wiley Sons, 2007.
55. Wiesemeyer, J. *New policy forces Canadian producers to cut back on fertilizer*. AgWeb. https://www.agweb.com/news/policy/politics/new-policy-forces-canadian-producers-cut-back-fertilizer, 2022.

56. Hoyer, S., Hamman, J., *xarray: N-D labeled arrays and datasets in Python*. Journal of Open Research Software. Ubiquity Press, vol. 5, no. 1, 2017.

A. Appendix

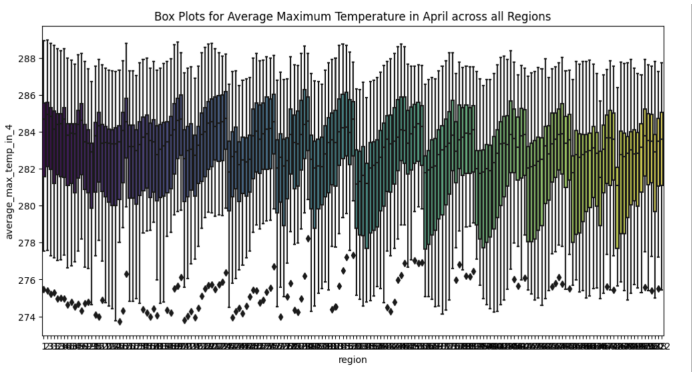


Figure 28. Boxplots for the Average Maximum Temperature in April across all Regions and Years.

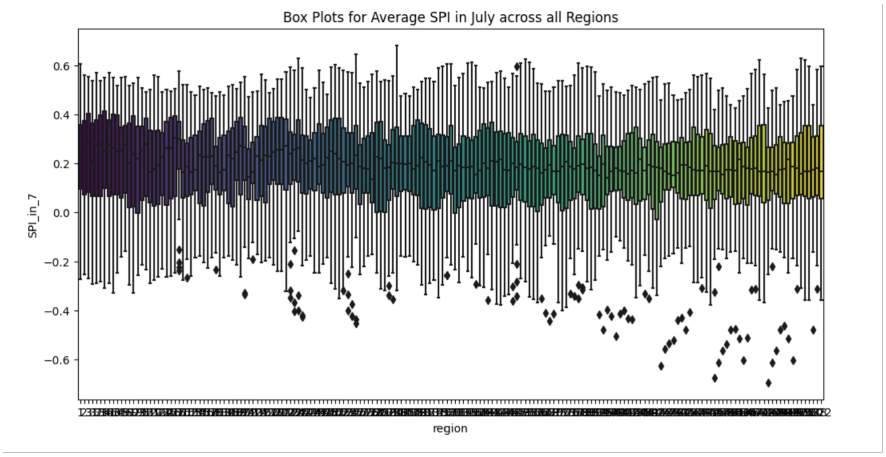


Figure 29. Boxplots for the SPI in July across all Regions and Years.

Category	Column	Description	Range	Unique Values
Index	time	Year	[1990, 2022]	33
Target Variable	Canola	standardized average Canola yield/ha	[-28.5, 37.3]	5973
Geographical Variables	latitude	Latitude coordinates	[49.0, 53.0]	41
	longitude	Longitude coordinates	[-110.0, -101.3]	88
	region	Region identifier	[1, 622]	181
Meteorological Variables	average_max_temp_in_4	Average maximum temperature in April	[273.74, 288.99]	5818
	average_max_temp_in_5	Average maximum temperature in May	[283.85, 296.17]	5786
	average_max_temp_in_6	Average maximum temperature in June	[288.91, 300.13]	5787
	average_max_temp_in_7	Average maximum temperature in July	[291.97, 303.74]	5763
	average_max_temp_in_8	Average maximum temperature in August	[290.06, 302.57]	5806
	average_max_temp_in_9	Average maximum temperature in September	[283.44, 298.98]	5808
	average_max_temp_in_10	Average maximum temperature in October	[274.65, 288.55]	5804
	days_above_25	Days above 25°C	[1, 95]	84
	days_under_0	Days under 0°C	[0, 0]	26
	longest_dry_spell	Length of longest dry spell	[0, 120]	22
	longest_wet_spell	Length of longest wet spell	[0, 120]	44
	days_over_95_precipitation	Days with over 95mm precipitation	[0, 120]	32
	longest_heat_wave	Length of longest heat wave	[0, 120]	14
	longest_cold_wave	Length of longest cold wave	[0, 120]	15
	SPI_in_4	Standardized Precipitation Index (SPI) in April	[-0.49, 0.78]	5379
	SPI_in_5	Standardized Precipitation Index (SPI) in May	[-0.54, 0.82]	5379
	SPI_in_6	Standardized Precipitation Index (SPI) in June	[-0.47, 0.64]	5379
	SPI_in_7	Standardized Precipitation Index (SPI) in July	[-0.69, 0.68]	5379
	SPI_in_8	Standardized Precipitation Index (SPI) in August	[-0.49, 0.87]	5379
	SPI_in_9	Standardized Precipitation Index (SPI) in September	[-0.44, 1.05]	5379
	SPI_in_10	Standardized Precipitation Index (SPI) in October	[-0.44, 0.92]	5379

Table 5. Description of the *canola_features* DataFrame.

Feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
avg_max_temp_4	-0.0303	-0.4914	0.2211	0.1638	-0.1547	0.0643	-0.0798	-0.0983	-0.0005	-0.2372
avg_max_temp_5	0.2714	-0.0111	0.3147	0.3402	0.1247	-0.0837	0.0092	0.2353	0.1271	0.1927
avg_max_temp_7	0.4090	-0.1151	-0.0351	-0.0819	-0.1515	0.2888	-0.2022	0.0045	0.0586	-0.0076
avg_max_temp_9	0.3074	-0.2059	-0.2749	-0.0712	0.3771	-0.1260	0.0496	0.1009	-0.1852	-0.0178
days_above_25	0.4875	-0.0700	-0.0122	0.0367	-0.0311	0.1648	0.0467	0.0521	0.1237	-0.0839
days_under_0	0.0524	0.5130	-0.0520	-0.0159	0.1343	0.3351	0.0016	-0.0178	0.0982	-0.0947
dry_spell	0.0652	-0.1151	0.0979	-0.3842	-0.0724	-0.1138	0.2743	-0.0927	0.5864	0.0887
wet_spell	-0.2719	-0.1163	-0.0400	0.1892	0.1949	0.0955	-0.1725	0.2687	0.4987	0.1855
heat_wave	0.2267	0.1862	-0.0366	-0.1028	-0.0662	-0.1070	-0.2592	-0.3546	0.2971	0.0835
cold_wave	-0.0354	0.3816	0.1239	0.1908	0.2583	0.2543	0.0995	-0.0300	-0.1664	-0.1077
SPI_in_4	-0.0797	0.1406	-0.2285	-0.3900	-0.1356	0.3204	0.2458	0.1539	0.0731	0.2246
SPI_in_5	-0.2095	-0.2230	-0.0923	-0.2677	-0.0563	0.2944	-0.0456	-0.2169	-0.0739	-0.4625
SPI_in_6	-0.2378	-0.2227	-0.1569	-0.0200	0.1407	0.3332	-0.1946	0.4661	0.0682	0.0431
SPI_in_7	-0.2863	-0.0326	0.0113	0.1692	0.1441	-0.1628	0.5652	-0.1632	0.1475	-0.0918
SPI_in_8	-0.2070	-0.1232	-0.1085	0.0263	0.1048	0.0883	-0.2794	-0.4880	-0.1542	0.6171
SPI_in_9	-0.1313	0.0958	0.4250	0.2205	-0.3877	0.3172	-0.0549	-0.0867	0.1040	-0.0006
SPI_in_10	0.0124	-0.0223	-0.1068	0.1510	0.4880	0.1137	-0.1830	-0.3315	0.3353	-0.3143
soil_zone	0.1583	-0.2390	0.2113	-0.0166	0.2269	0.3787	0.3650	-0.1673	-0.1268	0.2465
Longitude	-0.0123	0.0780	-0.5092	0.3607	-0.3356	-0.0763	0.0004	-0.0220	0.1048	-0.0651
Latitude	-0.1347	0.1393	0.3891	-0.3979	0.1904	-0.2336	-0.3135	0.1248	-0.0363	-0.1010

Table 6. Principal Components of Features

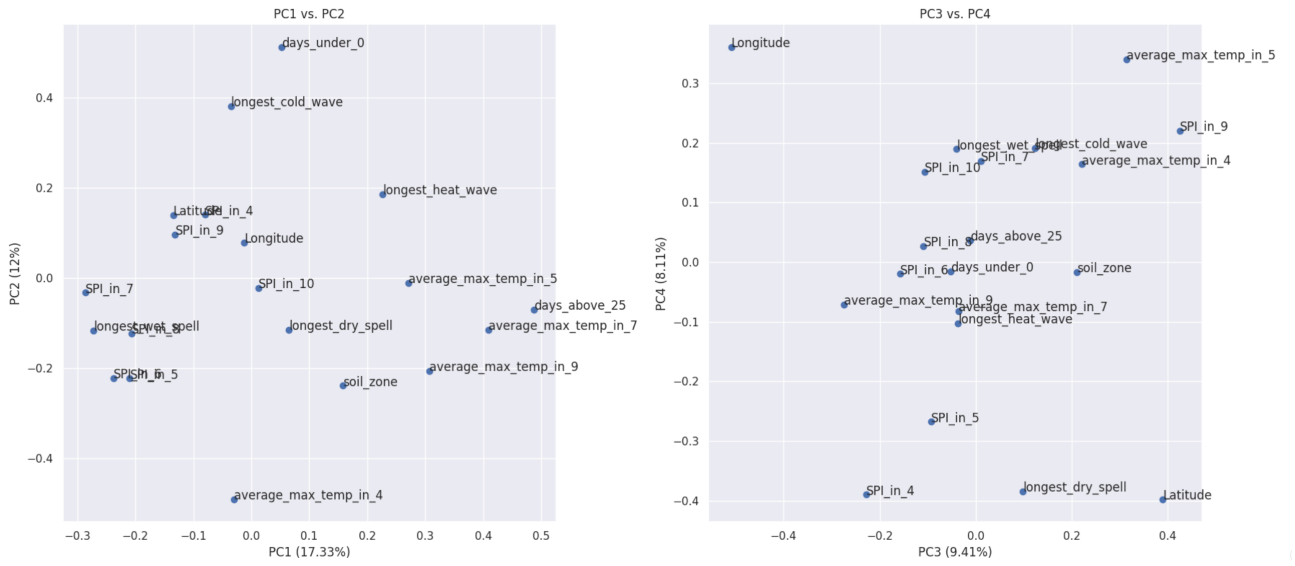


Figure 30. Variable Mappings in PC1 vs. PC2 and PC3 vs. PC4.

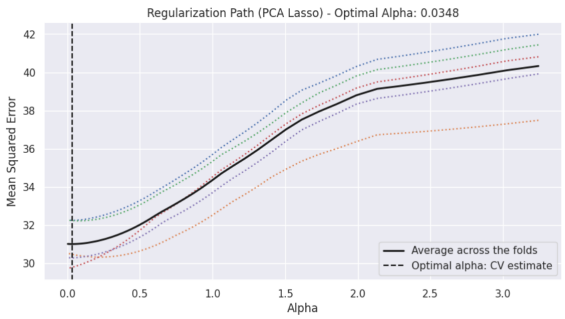


Figure 31. Determining an optimal value of *alpha* using 5-fold cross validation.

Feature_construction

March 15, 2024

1 Construction of the Feature Data Frame

[1]:
```
import pandas as pd
import numpy as np

from datetime import datetime
import glob
import xarray as xr
import os
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import statsmodels.api as sm

from statsmodels.tsa.holtwinters import SimpleExpSmoothing, ExponentialSmoothing

from sklearn.linear_model import LinearRegression

from sktime.transformations.series.detrend import Detrender
from sktime.forecasting.trend import PolynomialTrendForecaster
from sktime.utils.plotting import plot_series

from scipy.stats import gamma

from standard_precip.spi import SPI
from standard_precip.utils import plot_index

import seaborn as sns
```

1.1 Processing of the Weather Data

[2]:
```
#read dataframe
canola_2 = df = pd.read_csv('../data/rm-yields-data.csv', header=0,␣
 ↪index_col=0, parse_dates=True)
canola_small = canola_2.iloc[:, [0, 2]].copy()
```

[3]:
```
start_year = 1938
start_analysis = 1990
exclude_years = start_analysis - start_year

#cut of first 52 observations and 70s and 80s
canola_small.drop(canola_small.index[:exclude_years], inplace=True)

#filter out every observation that contains NAs
canola_filtered = canola_small.groupby('RM').filter(lambda group: not␣
 ↪group['Canola'].isnull().any())

# how may districts? 148
num_districts = canola_filtered.groupby('RM').ngroups
print(num_districts)
```

184

[4]:
```
# Group by 'RM' and check if 'Canola' has any missing values in each group
districts_with_full_data = canola_filtered.groupby('RM')['Canola'].apply(lambda␣
 ↪group: not group.isnull().any())

# Extract the list of districts with full data
districts_with_full_data_list =␣
 ↪districts_with_full_data[districts_with_full_data].index.tolist()
```

1.2 Processing of the Weather Data

[5]:
```
# select weather data

# open only the years from 1990 til 2022

# Define the directory path and pattern for the NetCDF files
directory_path = '../data/all_raw_data/'
file_pattern = 'data_*.nc'


# Get a list of files matching the pattern
files_to_open = glob.glob(os.path.join(directory_path, file_pattern))

# Open only the files for the years 1990 to 2022
years_to_open = list(map(str, range(start_analysis, 2023)))
files_to_open = [file for file in files_to_open if any(year in file for year in␣
 ↪years_to_open)]

# Use open_mfdataset to open the selected files
cop_all_90 = xr.open_mfdataset(files_to_open, combine='by_coords')
```

```
# open data from 1971 til 1989 as training data

training_years_to_open = list(map(str, range(1971, 1989)))
training_files_to_open = [file for file in files_to_open if any(year in file␣
 ↪for year in years_to_open)]

# Use open_mfdataset to open the selected files
training_data = xr.open_mfdataset(training_files_to_open, combine='by_coords')
```

[6]:
```
# center points for regions
df_regions = pd.read_csv(r'../data/cgn_sk_csv_eng.csv')
df_rms = df_regions[['Geographical Name','Latitude',␣
 ↪'Longitude']][df_regions['Generic Term'] == 'Rural Municipality']
df_rms['region_index'] = df_rms['Geographical Name'].str.split(' ').str[-1].
 ↪astype(int)
```

[7]:
```
def get_center(region):
    avg_lat = df_rms['Latitude'][df_rms['region_index'] == region].item()
    avg_long = df_rms['Longitude'][df_rms['region_index'] == region].item()
    return avg_lat, avg_long

def detrend_ts(df_region):
    # linear detrending
    forecaster = PolynomialTrendForecaster(degree=2)
    transformer = Detrender(forecaster=forecaster)
    yt = transformer.fit_transform(df_region['Canola'])
    return yt
```

[8]:
```
def merge_canola_weather_data(region = 310):
    # select data from region with center point
    center_lat, center_long = get_center(region)
    cropped_data_tmp = cop_all_90.sel(longitude=center_long,␣
 ↪latitude=center_lat,method='nearest')
    cropped_data_tmp_train = training_data.sel(longitude=center_long,␣
 ↪latitude=center_lat,method='nearest')

    # get residuals for canola yield
    df_tmp = canola_filtered[canola_filtered['RM'] == region]
    residuals = detrend_ts(df_tmp)

    # merge weather data and canola residuals
    df_weather_region = cropped_data_tmp.to_dataframe()
    df_weather_region_train = cropped_data_tmp_train.to_dataframe()

    df_weather_region['region'] = region
```

```
    column_to_append = residuals.tolist()
    years = df_weather_region.index.year
    df_weather_region['Canola_detrended'] = [column_to_append[year -␣
↪start_analysis] for year in years]
    df_weather_region.drop(['longitude','latitude'],axis=1,inplace=True)

    df_weather_region_train['region'] = region

    column_to_append = residuals.tolist()
    years = df_weather_region_train.index.year
    df_weather_region_train['Canola_detrended'] = [column_to_append[year -␣
↪start_analysis] for year in years]
    df_weather_region_train.drop(['longitude','latitude'],axis=1,inplace=True)

    return df_weather_region, pd.DataFrame(residuals), df_weather_region_train
```

1.3 Calculation of the Features

[9]:
```
def calculate_spi(prcp_data, scale=1):

    # Step 1: Calculate L-moments
    n = len(prcp_data)
    prcp_data_sorted = np.sort(prcp_data)

    # L-moment ratio
    l_moment_1 = np.sum(prcp_data_sorted) / n
    l_moment_2 = np.sum((2 * np.arange(1, n + 1) - 1 - n) * prcp_data_sorted) /␣
↪(n ** 2)

    # Step 2: Estimate parameters of gamma distribution
    k = l_moment_1 / l_moment_2
    theta = l_moment_2 / k

    # Step 3: Calculate SPI values
    spi_values = gamma.ppf((np.arange(1, n + 1) - 0.35) / (n + 0.3), a=k,␣
↪scale=theta * scale)

    return spi_values
```

[10]:
```
def calc_temp_features(df_weather_region, df_year):
    for month in range(4,11):
        daily_max_temperatures = df_weather_region.resample('D').max()
        monthly_avg_max_temperatures = daily_max_temperatures.resample('MS').
↪mean()

    #     dist1_df_month = dist1_df.resample('MS').mean()
```

```
        month_data = monthly_avg_max_temperatures[monthly_avg_max_temperatures.
↪index.month == month]
        column_to_append = month_data['t2m'].tolist()
        df_year.loc[:, f'average_max_temp_in_{month}'] = column_to_append
    return df_year
```

[11]:
```
def calc_spi_features(df_weather_region, df_year):
    for month in range(4, 11):
        # tried resampling in various ways but none worked
        tp_in_month = df_weather_region[df_weather_region.index.month == month]

        spi = SPI()

        # Assuming spi.calculate is the SPI calculation function
        spi_values = spi.calculate(
            tp_in_month.reset_index(),
            'time',
            'tp',
            freq="M",
            scale=1,
            fit_type="lmom",
            dist_type="gam"
        )

        # Add each SPI column separately
        for col_name in spi_values.columns:
            df_year[f'SPI_in_{month}_{col_name}'] = spi_values[col_name]

    return df_year
```

[12]:
```
def calc_summer_days(df_weather_region, df_year):

    testing_data = df_weather_region['t2m'].resample('D').max()
    testing_data = testing_data.dropna()
    test_df_summer = testing_data.to_frame(name='t2m')
    test_df_summer['month_day'] = test_df_summer.index.strftime('%m-%d')
    test_df_summer['above_25'] = test_df_summer['t2m'] > 298
    days_over_25 = test_df_summer.groupby(test_df_summer.index.
↪year)['above_25'].apply(sum)
    column_to_append = days_over_25.tolist()
    df_year.loc[:, f'days_above_25'] = column_to_append

    return df_year
```

[13]:
```
def calc_frost_days(df_weather_region, df_year):

    testing_data = df_weather_region['t2m'].resample('D').max()
```

```
    testing_data = testing_data.dropna()
    test_df_frost = testing_data.to_frame(name='t2m')
    test_df_frost['month_day'] = test_df_frost.index.strftime('%m-%d')
    test_df_frost['under_0'] = test_df_frost['t2m'] < 273
    days_under_0 = test_df_frost.groupby(test_df_frost.index.year)['under_0'].
↪apply(sum)
    column_to_append = days_under_0.tolist()
    df_year.loc[:, f'days_under_0'] = column_to_append

    return df_year
```

[14]:
```
# function to calculate the longest consecutive true streak

def longest_consecutive_true_streak(series):


    as_ints = series.astype(int)
    diff = as_ints.diff()
    groups = (diff == 1).cumsum()
    streak_lengths = as_ints.groupby(groups).sum()

    # Return the length of the longest streak
    return streak_lengths.max()
```

[15]:
```
def longest_dry_spell(df_weather_region, df_year):

    dist1_df_perci = df_weather_region.resample('D').sum()
    dist1_df_perci_wo0 = dist1_df_perci[dist1_df_perci['tp'] != 0].copy()

    # Threshhold is 1mm or 0.001m, here precipitation is measured in m

    dist1_df_perci_wo0.loc[:,'less_than_0.001'] = dist1_df_perci_wo0['tp'] < 0.
↪001
    longest_dry_spell_per_year = dist1_df_perci_wo0.groupby(dist1_df_perci_wo0.
↪index.year)['less_than_0.001'].apply(longest_consecutive_true_streak)

    column_to_append = longest_dry_spell_per_year.tolist()
    df_year.loc[:, f'longest_dry_spell'] = column_to_append

    return df_year
```

[16]:
```
def longest_wet_spell(df_weather_region, df_year):

    dist1_df_perci = df_weather_region.resample('D').sum()
    dist1_df_perci_wo0 = dist1_df_perci[dist1_df_perci['tp'] != 0].copy()

    # Threshhold is 1mm or 0.001m, here precipitation is likely measured in m
```

```
    dist1_df_perci_wo0.loc[:,'more_than_0.001'] = dist1_df_perci_wo0['tp'] > 0.
↪001
    longest_wet_spell_per_year = dist1_df_perci_wo0.groupby(dist1_df_perci_wo0.
↪index.year)['more_than_0.001'].apply(longest_consecutive_true_streak)

    column_to_append = longest_wet_spell_per_year.tolist()
    df_year.loc[:, f'longest_wet_spell'] = column_to_append

    return df_year
```

[17]:
```
def over_95_precipitation(df_weather_region, df_weather_region_train, df_year):

    dist1_df_perci = df_weather_region.resample('D').sum()
    dist1_df_perci_wo0 = dist1_df_perci[dist1_df_perci['tp'] != 0]
    testing_data_pre = dist1_df_perci_wo0["tp"]

    dist1_df_perci_train = df_weather_region_train.resample('D').sum()
    dist1_df_perci_wo0_train = dist1_df_perci_train[dist1_df_perci_train['tp'] !
↪= 0]
    training_data_pre = dist1_df_perci_wo0_train["tp"]


    # calculate for every day the 90% quantile
    quantile_95_series = training_data_pre.groupby([training_data_pre.index.
↪month, training_data_pre.index.day]).quantile(0.95)
    quantile_95_series.index = quantile_95_series.index.map(lambda x: f"{x[0]:
↪02d}-{x[1]:02d}")

    test_df_pre = testing_data_pre.to_frame(name='tp')
    test_df_pre['month_day'] = test_df_pre.index.strftime('%m-%d')

    # map the 90th percentile values from quantile_90_series to the test series
    test_df_pre['quantile_95'] = test_df_pre['month_day'].apply(lambda x:␣
↪quantile_95_series.get(x, pd.NA))

    # compare each test value to its corresponding 90th percentile value
    test_df_pre['over_quantile_95'] = test_df_pre['tp'] >␣
↪test_df_pre['quantile_95']
    test_df_pre.drop(['month_day', 'quantile_95'], axis=1, inplace=True)

    # Group the DataFrame by year, and apply the function to find the longest␣
↪streak of True values
    days_over_95 = test_df_pre.groupby(test_df_pre.index.
↪year)['over_quantile_95'].apply(sum)
```

```
    column_to_append = days_over_95.tolist()
    df_year.loc[:, f'days_over_95_precipitation'] = column_to_append

    return df_year
```

[18]:
```
def heat_wave(df_weather_region, df_weather_region_train, df_year):

    training_data = df_weather_region_train['t2m'].resample('D').max()
    training_data = training_data.dropna()

    testing_data = df_weather_region['t2m'].resample('D').max()
    testing_data = testing_data.dropna()

    # calculate for every day the 90% quantile
    quantile_90_series = training_data.groupby([training_data.index.month,␣
↪training_data.index.day]).quantile(0.9)
    quantile_90_series.index = quantile_90_series.index.map(lambda x: f"{x[0]:
↪02d}-{x[1]:02d}")

    test_df = testing_data.to_frame(name='value')
    test_df['month_day'] = test_df.index.strftime('%m-%d')

    # map the 90th percentile values from quantile_90_series to the test series
    test_df['quantile_90'] = test_df['month_day'].apply(lambda x:␣
↪quantile_90_series.get(x, pd.NA))

    # compare each test value to its corresponding 90th percentile value
    test_df['is_above_quantile_90'] = test_df['value'] > test_df['quantile_90']
    test_df.drop(['month_day', 'quantile_90'], axis=1, inplace=True)

    # Group the DataFrame by year, and apply the function to find the longest␣
↪streak of True values
    longest_heat_streak_by_year = test_df.groupby(test_df.index.
↪year)['is_above_quantile_90'].apply(longest_consecutive_true_streak)

    column_to_append = longest_heat_streak_by_year.tolist()
    df_year.loc[:, f'longest_heat_wave'] = column_to_append

    return df_year
```

[19]:
```
def cold_wave(df_weather_region, df_weather_region_train, df_year):

    training_data = df_weather_region_train['t2m'].resample('D').min()
    training_data = training_data.dropna()

    testing_data = df_weather_region['t2m'].resample('D').min()
    testing_data = testing_data.dropna()
```

```
    # calculate for every day the 90% quantile
    quantile_10_series = training_data.groupby([training_data.index.month,␣
↪training_data.index.day]).quantile(0.1)
    quantile_10_series.index = quantile_10_series.index.map(lambda x: f"{x[0]:
↪02d}-{x[1]:02d}")

    test_df_cold = testing_data.to_frame(name='value')
    test_df_cold['month_day'] = test_df_cold.index.strftime('%m-%d')

    # map the 90th percentile values from quantile_90_series to the test series
    test_df_cold['quantile_10'] = test_df_cold['month_day'].apply(lambda x:␣
↪quantile_10_series.get(x, pd.NA))

    # compare each test value to its corresponding 90th percentile value
    test_df_cold['is_under_quantile_10'] = test_df_cold['value'] <␣
↪test_df_cold['quantile_10']
    test_df_cold.drop(['month_day', 'quantile_10'], axis=1, inplace=True)

    # Group the DataFrame by year, and apply the function to find the longest␣
↪streak of True values
    longest_streak_by_year_cold = test_df_cold.groupby(test_df_cold.index.
↪year)['is_under_quantile_10'].apply(longest_consecutive_true_streak)

    column_to_append = longest_streak_by_year_cold.tolist()
    df_year.loc[:, f'longest_cold_wave'] = column_to_append

    return df_year
```

[20]:
```
available_regions = [region for region in districts_with_full_data_list if␣
↪region in df_rms['region_index'].to_list()]
```

[21]:
```
# remove problematic regions

available_regions.remove(278)
available_regions.remove(529)
```

1.4 Construction of the final Data Frame

[22]:
```
dfs_of_years = []
for region in available_regions:
    df_weather_region, df_year, df_weather_region_train  =␣
↪merge_canola_weather_data(region)
    df_year.index = df_year.index.year
    df_year['region'] = region
    df_year = calc_temp_features(df_weather_region, df_year)
    df_year = calc_summer_days(df_weather_region,df_year)
```

```
    df_year = calc_frost_days(df_weather_region,df_year)
    df_year = longest_dry_spell(df_weather_region,df_year)
    df_year = longest_wet_spell(df_weather_region,df_year)
    df_year = over_95_precipitation(df_weather_region, df_weather_region_train,␣
 ↪df_year)
    df_year = heat_wave(df_weather_region, df_weather_region_train, df_year)
    df_year = cold_wave(df_weather_region, df_weather_region_train, df_year)
    dfs_of_years.append(df_year)
```

[23]:
```
df_full = pd.concat(dfs_of_years)
```

[28]:
```
files_dir = '../data/all-spi-features'
```

[26]:
```
files = glob.glob(os.path.join(files_dir, "*.csv"))

dataframes = [pd.read_csv(file) for file in files]

spis_df = pd.concat(dataframes, ignore_index=False)

spis_df.set_index('Year', inplace=True)
```

[29]:
```
merged_df = pd.merge(df_full, spis_df, how='left', left_on=['Year', 'region'],␣
 ↪right_on=['Year', 'Region'])
merged_df.head()
```

[29]:

Year	Canola	region	average_max_temp_in_4	average_max_temp_in_5	\
1990	0.127132	1	284.965759	290.526489	
1991	2.520378	1	287.648346	291.984650	
1992	-6.339489	1	283.242584	293.701874	
1993	4.147971	1	284.697113	292.762695	
1994	2.081733	1	285.144440	293.432678	

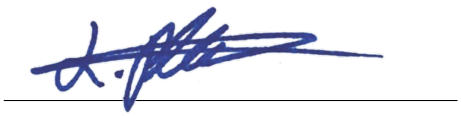
Year	average_max_temp_in_6	average_max_temp_in_7	average_max_temp_in_8	\
1990	297.082458	299.412781	300.752075	
1991	296.854401	297.652344	299.938751	
1992	295.445709	294.548126	297.033264	
1993	293.425385	294.037415	296.177429	
1994	294.774536	297.061951	297.072021	

Year	average_max_temp_in_9	average_max_temp_in_10	days_above_25	…	\
1990	296.791382	286.036377	72	…	
1991	292.543549	282.321960	51	…	
1992	292.229004	285.503967	37	…	
1993	291.031281	284.688232	26	…	

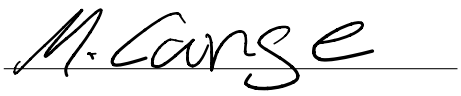
```
1994             294.903839              286.052094             38  …

      longest_cold_wave  SPI_in_4  SPI_in_5  SPI_in_6  SPI_in_7  SPI_in_8  \
Year
1990                  5  0.091579  0.029108  0.115236  0.209242  0.256968
1991                  4  0.331210  0.331767  0.195163  0.282696  0.222439
1992                  7  0.144230  0.193160 -0.188189  0.358708  0.309846
1993                  5  0.115240 -0.052129  0.086120  0.608185  0.263007
1994                  6 -0.063990  0.114090  0.446020  0.102491  0.320304

      SPI_in_9  SPI_in_10  Region  Soil Zone
Year
1990  0.171961  -0.032910       1          1
1991  0.398697   0.232995       1          1
1992  0.225539  -0.114548       1          1
1993  0.235375  -0.044934       1          1
1994  0.034789   0.322164       1          1

[5 rows x 25 columns]
```

B. Declaration of Autonomy

We hereby declare that we wrote this thesis paper independently, without assistance from external parties, and without use of resources other than those indicated. All information taken from other publications or sources in text or in meaning are duly acknowledged in the text. The written and electronic forms of the thesis paper are the same. We give our consent to have this thesis checked by plagiarism software.



Laura Plodek
Göttingen, March 15th, 2024



Marisa Lange
Göttingen, March 15th, 2024