

# Data-driven and machine learning framework for Alfalfa yield response to long-term climate variability in the Kansas High Plains (USA)

Received: 3 July 2025

Accepted: 25 May 2026

Published online: 04 June 2026

Cite this article as: Nazrul F., Kim J., Dey S. *et al.* Data-driven and machine learning framework for Alfalfa yield response to long-term climate variability in the Kansas High Plains (USA). *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-55545-2>

Farshina Nazrul, Jiyung Kim, Sourajit Dey, Bradley Whitaker, Debjani Sihi, Doohong Min & Gaurav Jha

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

# Data-Driven and Machine Learning Framework for Alfalfa Yield Response to Long-Term Climate Variability in the Kansas High Plains (USA)

## **Authors**

Farshina Nazrul<sup>1</sup>, Jiyung Kim<sup>2</sup>, Sourajit Dey<sup>2</sup>, Bradley Whitaker<sup>1</sup>, Debjani Sihi<sup>3</sup>, Doohong Min<sup>2</sup>, Gaurav Jha<sup>2\*</sup>

## **Affiliations**

<sup>1</sup>Department of Electrical & Computer Engineering, Montana State University, Bozeman, MT 59717, USA

<sup>2</sup>Department of Agronomy, Kansas State University, Manhattan, KS 66506, USA

<sup>3</sup>Department of Plant and Microbial Biology, Department of Crop and Soil Sciences, N.C. Plant Sciences Initiative, North Carolina State University, Raleigh, NC 27695, USA

\*Corresponding author. Email: gjha@ksu.edu

## **Abbreviations:**

USDA National Agricultural Statistics Service (USDA-NASS), Parameter elevation Regression on Independent Slopes Model (PRISM), Relative Humidity (RH), Vapor Pressure Deficit (VPD), Growing Degree Days (GDD), Machine Learning (ML), Minimum Redundancy Maximum Relevance (mRMR), Mutual Information (MI), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Cross Validation (CV), Root Mean Squared Error (RMSE)

## **ABSTRACT**

Climate extremes and declining water availability in the U.S. High Plains threaten the long-term sustainability of alfalfa production. This study presents a machine learning (ML) based modeling framework to identify and evaluate key agro-climatic predictors of alfalfa (*Medicago sativa* L.) yield across nine agricultural districts in Kansas from 1981 to 2018. A novel contribution of this study is the integration of a high-dimensional climate predictor space, long-term historical records, and district-level stratification of irrigated and rainfed systems to systematically assess spatially varying climate–yield relationships. Using a high-dimensional feature space of 117 climate-derived variables from PRISM and yield data from USDA-NASS, a Minimum Redundancy Maximum Relevance (mRMR) algorithm was applied to rank predictors, followed by forward feature selection with three non-parametric regression models. Four spatial configurations were evaluated—statewide, district-specific, irrigated, and rainfed datasets. The Kansas High Plains region, with its strong west-to-east precipitation gradient, groundwater-dependent irrigation systems, and recurrent drought exposure, offers a globally relevant testbed for understanding crop-climate interactions in semi-arid environments. Model evaluation RMSE during feature selection ranged from 0.52 to 1.14 tons/acre. Dew point temperature in August and seasonal VPD metrics emerged as dominant predictors, indicating these variables are strongly associated with late-summer

yield variability. Rainfed districts prioritized drought indicators (e.g., no precipitation days, VPD), while irrigated districts showed stronger associations with thermal and humidity metrics. By explicitly accounting for climate sensitivities across management regimes and spatial scales, this framework provides a transferable approach for identifying robust, region-specific climate drivers in high-dimensional agricultural datasets.

**Keywords: Alfalfa, Climate Variables, Feature Selection, Minimum Redundancy Maximum Relevance (mRMR) algorithm, Machine Learning Framework, Kansas High Plains**

## 1 INTRODUCTION

Alfalfa (*Medicago sativa* L.) is a perennial forage legume with high forage yield<sup>1,2</sup>, high nutritive value<sup>1,2</sup>, and biological nitrogen fixation<sup>2,3</sup> of more than 80% of the amount required<sup>4</sup>. Kansas ranks as the 9<sup>th</sup> largest alfalfa-producing state in the USA in 2023, where annual alfalfa hay production exceeded 2.2 million tons from about 1.73 million acres<sup>5</sup>.

The sustainability of alfalfa production in the Kansas High Plains is under increasing pressure due to intensified climate extremes (severe to exceptional droughts) and long-term declines in groundwater levels from the Ogallala Aquifer<sup>6</sup>. The water levels have declined by 25 to 103 feet across five groundwater management districts, leading to almost a 60% reduction of the original water level in some areas since predevelopment in the 1950s<sup>7</sup>. The critical role of groundwater, particularly from the Ogallala Aquifer, was emphasized in sustaining crop production across the Great Plains<sup>8</sup>. Alfalfa yields can drop over 70% during drought<sup>9,10</sup>. As a result, the anticipated decline in the level of the Ogallala aquifer poses a future challenge for alfalfa cultivation in Kansas<sup>11</sup>. The amount and timing of precipitation during the growing season are critical for alfalfa growth, regrowth, and total biomass accumulation.

Alfalfa is typically established in late summer or early fall to reduce weed competition and favorable soil moisture in Kansas<sup>12</sup>. Once established, alfalfa can be harvested multiple times per growing season, often beginning in May and extending through the last week of September or the first fortnight of October, depending on precipitation and irrigation availability<sup>13</sup>. In Kansas, the growing season window for alfalfa differs regionally. In western Kansas, warmer spring temperatures allow green-up to begin as early as early March, while in eastern Kansas, this typically occurs later in the month. Similarly, the final harvest can be extended until the last week of October in western Kansas, whereas in eastern Kansas, the last cutting is recommended no later than the first week of October to reduce the risk of winter injury<sup>13,14</sup>. First cutting is usually done at 10% flowering, followed by subsequent cuttings with a 28-32 days interval, balancing forage yield with nutritive value, particularly crude protein and digestibility. Alfalfa grows well in well-drained soil but requires a lot of water, with a water demand from 26 to 40 inches. In Kansas, where water availability varies substantially across regions, alfalfa serves as a critical forage crop for both dairy and beef production, and its productivity is highly sensitive to climatic variability<sup>11</sup>.

To better understand and manage alfalfa growth, several key climatic variables are often considered. Precipitation directly influences soil moisture availability and irrigation demand, while temperature regulates growth rate, dormancy, and seasonal development<sup>15</sup>. Vapor pressure deficit (VPD) serves as a critical indicator of atmospheric water demand, influencing evapotranspiration and plant water stress, whereas dew point temperature reflects ambient moisture conditions that affect transpiration efficiency<sup>16,17,18</sup>. A high dew point temperature causes the winterkill and summer slump in winter and summer, respectively. The VPD at a given temperature can affect plant growth due to its impact on stomatal aperture, carbon dioxide (CO<sub>2</sub>) uptake, transpiration, or leaf temperature<sup>19,20</sup>. It is a useful parameter to evaluate the drought condition<sup>21</sup> and water productivity<sup>22,23</sup>. The low

VPD negatively affects plant growth and productivity<sup>24</sup>. Alfalfa, being a perennial crop with distinct growth cycles influenced by varying weather conditions, particularly VPD, experiences fluctuations in water productivity<sup>25</sup>. The optimal temperature range for alfalfa growth typically falls between 65 to 85 °F<sup>1</sup>, with alfalfa capable of tolerating the maximum daily growth temperature of up to 100 °F<sup>26</sup>. Together, these variables provide an integrated view of water supply and atmospheric demand, which are central to alfalfa production in irrigated and rainfed environments of Kansas.

Given the impacts of climate variability, there is growing interest in using machine learning (ML) techniques<sup>27,28</sup>. Recent studies have applied ML approaches to evaluate climate-driven agricultural responses and water-use dynamics, demonstrating the value of data-driven models for capturing nonlinear relationships and future climate impacts<sup>29,30,31,32</sup>. In parallel, several studies have also explored spatial variability in water productivity and drought resilience in alfalfa and other cropping systems using field-based and high-throughput approaches, highlighting the complexity of climate–yield interactions across management conditions<sup>33,34</sup>.

ML approaches such as feature selection offer tools for handling high-dimensional climate data and uncovering complex interactions that traditional statistical models may overlook in crop model simulation and crop yield predictions<sup>29,30</sup>. Among various feature selection techniques, the Minimum Redundancy Maximum Relevance (mRMR) algorithm has shown potential in analyzing the complex, interrelated climate variables that show high correlation with crop yield. The mRMR selects features that are both highly relevant to the target outcome (i.e., yield) and minimally redundant with one another, making it well-suited for high-dimensional environmental data<sup>31,32</sup>. Its ability to balance relevance and redundancy enhances the accuracy and interpretability of the associated ML models in both classification and regression tasks<sup>33,34,35</sup>. Alfalfa is an important forage crop in the Kansas High Plains, where its productivity plays an important role in both agricultural systems and regional water use. However, knowledge of how long-term climate variability has shaped alfalfa yield trends in this region remains limited. Previous studies have largely focused on short-term field experiments, region-wide modeling using a small set of climate variables, or coarse-scale assessments that do not consider individual impact on the nine agricultural districts<sup>4,8,36,37,38,13,39</sup>. These approaches often overlook the high dimensionality and interdependence of variables in climate datasets that are critical to understanding crop-climate interactions over time<sup>27,28</sup>. Moreover, only a few studies have applied statistical and data-driven feature selection techniques, such as mRMR, to identify the most informative predictors of alfalfa yield across diverse agro-climatic settings.

To address these limitations, this study presents a high-dimensional analysis of long-term climate impacts on alfalfa yield in the Kansas High Plains. In this regard, a high-dimensional feature selection strategy was integrated with an ML-based regression framework to uncover the most relevant climate predictors of alfalfa yield across both irrigated and rainfed conditions in Kansas. By integrating 117 climate-derived predictors over a 38-year period (1981–2018) and differentiating irrigated vs. rainfed systems through the individual analysis of nine agricultural districts<sup>37,38,39,22</sup> and grouped subclasses according to the irrigation water management practices, this study addresses this key methodological and contextual gap using a data-driven framework grounded in yield prediction and climate relevance ranking.

The Kansas High Plains was chosen as the testbed for this study because it represents one of the most agriculturally productive yet environmentally vulnerable regions in North America, characterized by a steep west-to-east precipitation gradient, recurrent droughts, and long-term depletion of the Ogallala Aquifer, one of the largest freshwater aquifers in the world. This study aims to rank a broad set of climatic features, capturing both seasonal variance through monthly measurements and long-term variance through annual measurements, to reflect temporal and environmental variability across Kansas. The objective of this study is to also assess the predictive capability of these ranked variables

across irrigated and rainfed conditions within each agricultural district through ML-based regression models and forward selection. As water constraints intensify under climate change, this study offers a scalable ML framework to identify climate-sensitive yield drivers, providing insights that can directly inform and guide crop management decisions. For example, if precipitation and temperature extremes consistently emerge as top predictors, that information can help refine when and how much to irrigate as Ogallala water supplies decline. Seasonal patterns in climate drivers can also guide when fields should be cut, so harvests avoid the worst periods of heat or drought stress. Likewise, knowing which factors are most predictive across districts can support the choice of more drought-tolerant cultivars in drier areas, or help anticipate when supplemental forage might be needed in high-risk years.

## 2 MATERIALS AND METHODS

This study investigated the relationship between climate variables and alfalfa yield across Kansas using a data-driven feature selection and evaluation framework (Fig. 1). County-level yield data from USDA-NASS<sup>40</sup> and daily climate data from the PRISM Climate Group<sup>41</sup> were collected for the years 1981–2018. The time series data for each county was generated on Jul 06, 2023. These county series were derived from the 4 km resolution PRISM AN81d gridded data within each county boundary<sup>42</sup>.

From the climate data, 117 features were computed to represent monthly and annual temperature, precipitation, humidity, and other climatic variables. To account for geographic and hydrological variability in yield response, the analysis was performed across four types of datasets: (i) all-districts combined (statewide), (ii) nine agricultural districts individually<sup>37,38,39,22</sup>, (iii) districts with primarily irrigated conditions, and (iv) rainfed districts. This classification reflects historical differences in groundwater access, precipitation patterns, and irrigation practices across the state and allowed us to investigate both regional patterns and the role of water availability in climate-yield relationships<sup>43,44</sup>.

To rank the agro-climatic variables as per their importance to alfalfa yield, the mRMR algorithm<sup>31</sup> was implemented. Forward feature selection was then conducted by incrementally evaluating subsets of variables using three ML-based data-driven regression models: SVM<sup>45,46</sup>, KNN<sup>47,48</sup>, and RF<sup>49,50</sup>, respectively. Each model was trained and tested using 5-fold CV<sup>51</sup>, and performance was assessed using the regression RMSE<sup>52</sup>. For each dataset, the best observed feature set was selected based on the first local minimum in the moving average of the average RMSE across the three models. Fig. 1 presents an overview of the methodological workflow.

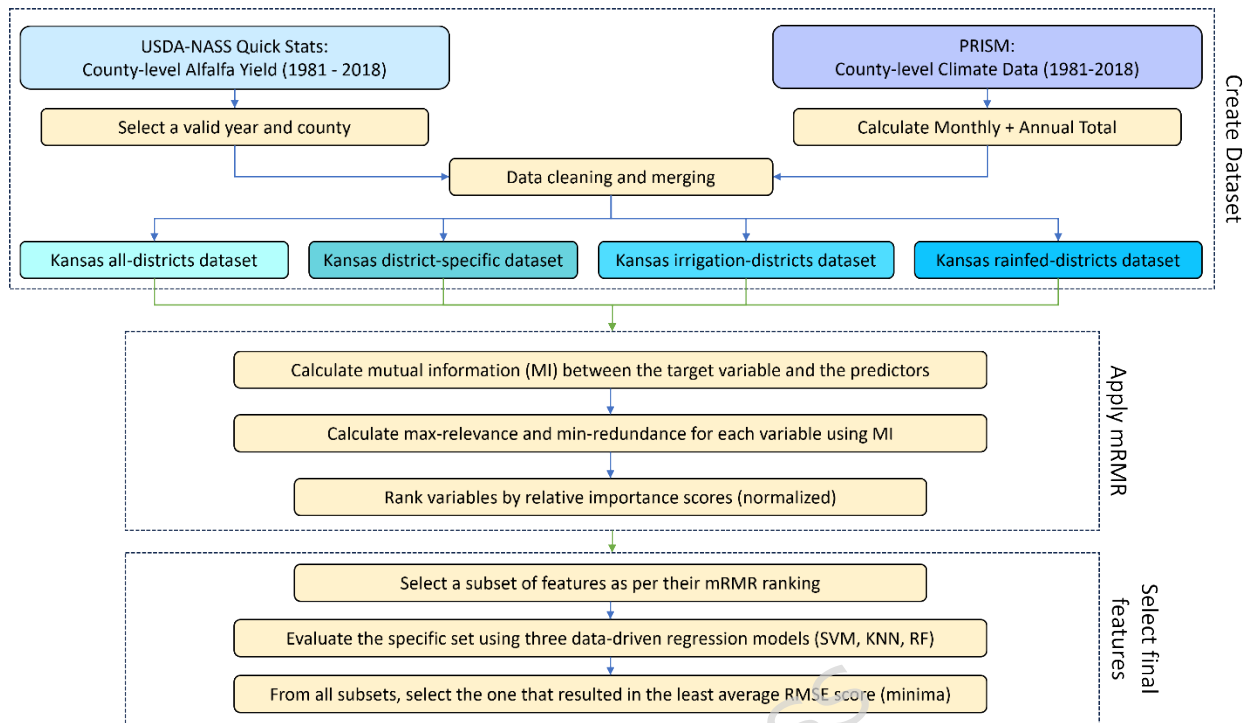


Fig. 1. Overview of the methodology for selecting climate features most relevant to the alfalfa yield in Kansas using mRMR and regression models (SVM, KNN, RF) across four dataset types: all-districts, district-specific, irrigated, and rainfed. The final feature set for each dataset corresponds to the subset with the lowest observed RMSE (based on 5-fold CV) during forward selection.

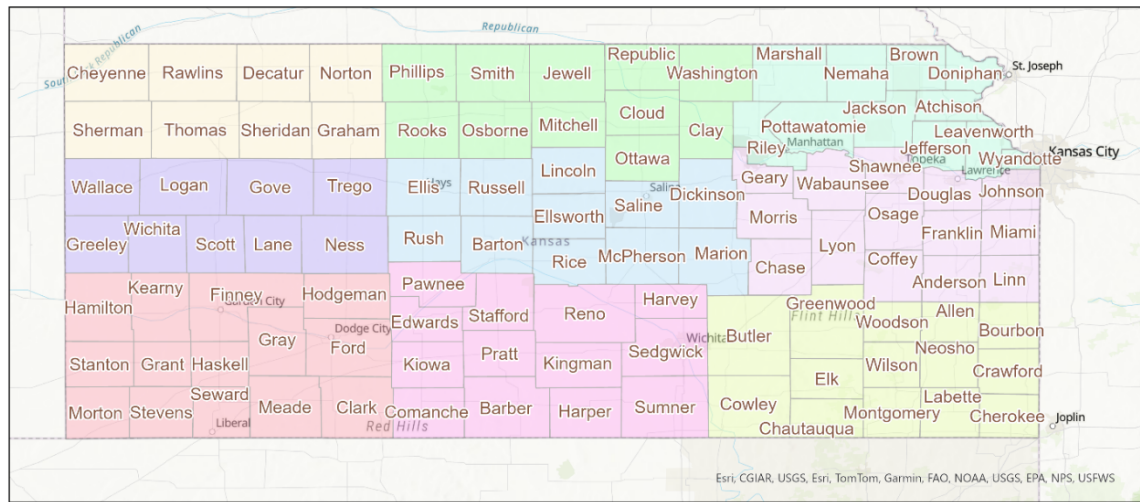
## 2.1 Location Description

Alfalfa production in Kansas is shaped by significant climatic and hydrological gradients, which are reflected in its division into nine agricultural districts<sup>37</sup>. These districts, established by the USDA, align with county boundaries and are designed to capture regional differences in climate, topography, and water resource availability, providing a structured framework for agricultural data collection and analysis<sup>38</sup>. Fig. 2 (a) shows these nine districts, with counties labeled and color-coded by district to highlight regional distinctions. The average annual precipitation for each county during the study period are shown in Fig. 2 (b).

The east-west precipitation gradient across Kansas, ranging from 20 inches in the semi-arid west to 40 inches in the humid east<sup>53</sup>, is a major factor in differentiating the nine agricultural districts. This gradient directly influences water availability during the alfalfa growing season (April-October), with western districts often facing moderate to severe droughts that limit rainfed alfalfa production. The precipitation gradient is mirrored by an elevation rise from approximately 650 feet in the east to over 3900 feet in the west, influencing both temperature regimes and growing conditions<sup>41</sup>. Additionally, a north-south temperature gradient exists, with southern Kansas experiencing slightly warmer temperatures (about 58 °F average annual temperature) compared to the cooler northern regions (about 52 °F average annual temperature), further influencing the spatial variability in alfalfa growth conditions<sup>54</sup>.

To address the water differences, the nine districts were categorized into two sub-groups based on water management practices (Fig. 2): six districts with primarily irrigated conditions (districts 10-60) and three with rainfed conditions (districts 70-90). This classification was informed by data on groundwater usage, irrigation infrastructure, and

precipitation patterns. In irrigated districts, access to groundwater resources, such as the Ogallala Aquifer, supports supplemental irrigation to mitigate drought stress. Conversely, rainfed districts rely solely on natural precipitation, making them more susceptible to yield reductions during dry periods<sup>43</sup>.



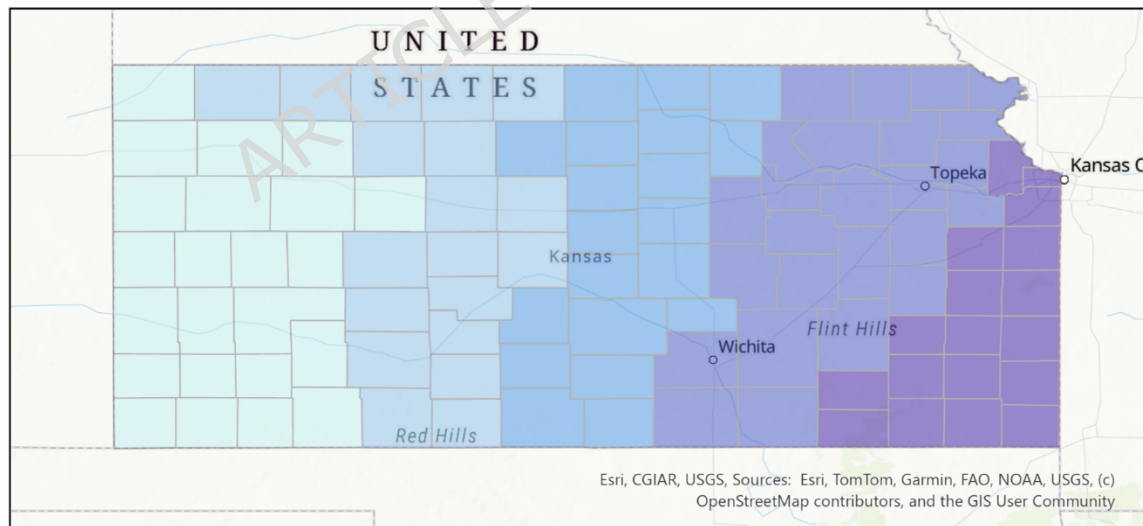
AgDistrictCode



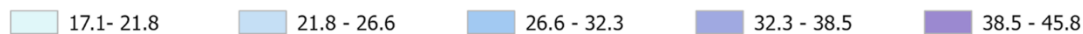
0 55 110 220 Miles



(a)



Average Annual Precipitation (inches)



(b)

Fig. 2. County-level map of Kansas showing (a) the nine agricultural districts color-coded by region. Based on water management practices, these nine districts were

classified into two categories: six irrigation districts (10–60) and three rainfed districts (70–90); (b) Average annual total precipitation per county during the study period (1981–2018).

## 2.2 Dataset Description

### 2.2.1 Data sources

In this study, publicly available annual yield data of alfalfa from Kansas were combined with the corresponding weather parameters. The county-level yield data (tons/acre) for the years 1981 to 2018 were obtained from USDA-NASS<sup>40</sup>. The relevant climatic information were obtained from PRISM; precipitation, the minimum, mean, and maximum temperatures, the mean dew point temperature, and both the minimum and maximum values of vapor pressure deficit<sup>41</sup>. County-level yield estimates from USDA-NASS are derived from survey-based data and are subject to uncertainties associated with sampling error and nonresponse bias during reporting variability. These uncertainties are quantified by NASS using standard errors and coefficients of variation, although the magnitude of uncertainty varies depending on crop, region, and sample size.

### 2.2.2 Preprocessing

**Observations:** The annual yield data from USDA-NASS<sup>40</sup> were available for each year and for all counties and reported in tons per acre. However, starting from 2004, some of the counties collectively reported their average yield (109 in total). As this study focused more on analyzing the county-specific agro-climatic relationship, the combined county reports were removed while creating the dataset. As a result, we finally obtained a total of 3353 (84%) data points out of a total of 3990 observations for 105 Kansas counties over 38 years. The detailed county-specific missing years are reported in Table S1 in the supplementary material.

**Features:** Seven daily weather parameters: precipitation (ppt), minimum temperature (tmin), average temperature (tmean), maximum temperature (tmax), average dew point temperature (tdmean), minimum vapor pressure deficit (vpdmin) and maximum vapor pressure deficit (vpdmax) from PRISM climate data<sup>41</sup> were extracted for each county. The daily values were then combined to calculate the monthly and annual total for each of the parameters for each county from 1981 to 2018. This resulted in 13 features (12 monthly + 1 annual) for each of the seven climatic parameters obtained from PRISM (91 features in total for each of the 3353 points in our dataset). In addition, two additional parameters were calculated (both on a monthly and annual basis)—days with no precipitation (nopptdays) and Growing Degree Days (GDD).

**Days with no precipitation (nopptdays):** The number of days with more than 0 mm precipitation can significantly influence alfalfa growth, particularly during critical stages of development<sup>15</sup>. Especially during hot and sunny days, no precipitation can increase evapotranspiration demand while reducing soil moisture, limiting alfalfa growth and yield if irrigation is not available<sup>55</sup>. To quantify dryness, the number of days with no precipitation were calculated, both on a monthly and annual basis. A binary indicator variable,  $\text{nopptday}_n = \{0,1\}$  was defined, where, for each day  $n$  in the month or a year, 0 indicates no precipitation (0 mm) and 1 indicates precipitation above 0 mm. Then, the total number of ‘nopptdays’ over a one-month or one-year period was computed as follows.

$$\text{nopptdays}_{\text{total}} = \sum_{n=1}^N \text{nopptday}_n \quad (1)$$

where,  $N$  is the total number of days in the specified period (either 28/30/31 for a month or 365/366 for a year), and  $n$  represents each calendar day within that period.

**Total number of Growing Degree Days (GDD<sub>total</sub>):** Typically, the growing season for alfalfa is defined as the days with average temperature greater than 5 °C (base temperature for alfalfa production)<sup>10,56</sup>. In this study, for each month and year, the

corresponding number of GDD was counted as the total number of days that resulted in a positive value for the following equation. Therefore, first, a binary indicator variable,  $GDD_n = \{0,1\}$  was defined for each day, based on whether the corresponding GDD for that day is a positive number.

$$GDD_n = \begin{cases} 1, & \frac{t_{max}+t_{min}}{2} - t_{base} \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $n$  represents each day between the years 1981 and 2018.

Then, the total number of days where  $GDD_n = 1$  was calculated over a monthly or annual period.

$$GDD_{total} = \sum_{n=1}^N GDD_n \quad (3)$$

where  $N$  is the total number of days in a month (for monthly GDD) or year (for annual GDD). Hence, in this study, including the 'nopptdays' and ' $GDD_{total}$ ' parameters in addition to the PRISM climatic features, we used a total of 117 agro-climatic features. Table 1 summarizes these 117 agro-climatic features used in this study, including their variable names, temporal scale, units, and total feature count.

In this study, GDD was used as a climate-derived indicator of thermal accumulation and broad seasonal development potential. However, GDD does not directly represent cultivar-specific phenology or genetic differences in alfalfa growth habit. Alfalfa cultivars differ in fall dormancy, winter hardiness, regrowth rate after cutting, root architecture, drought tolerance, heat tolerance, and persistence, all of which can influence yield response under similar climatic conditions. Because the USDA-NASS county-level yield records used in this study do not include cultivar identity, dormancy class, stand age, genetic background, or cutting schedule, genetic effects could not be explicitly modeled. Therefore, GDD should be interpreted as an environmental proxy for thermal opportunity rather than a direct measure of genotype-specific developmental response.

**Table 1. Summary of climate features used for agro-climatic feature ranking for alfalfa yield in Kansas (1981-2018).**

| Agro-climatic Feature Type | Variable Name(s)  | Data Scale | Unit   | No. of Total Features |
|----------------------------|-------------------|------------|--------|-----------------------|
| Precipitation              | ppt               | Monthly    | inches | 12                    |
|                            |                   | Annual     | inches | 1                     |
| Temperature                | tmin, tmean, tmax | Monthly    | °F     | 12×3=36               |
|                            |                   | Annual     | °F     | 1×3=3                 |
| Dew Point Temperature      | tdmean            | Monthly    | °F     | 12                    |
|                            |                   | Annual     | °F     | 1                     |
| Vapor Pressure Deficit     | vpdmin, vpdmax    | Monthly    | hPa    | 12×2=24               |
|                            |                   | Annual     | hPa    | 1×2=2                 |
| Days with no Precipitation | nopptdays         | Monthly    | days   | 12                    |
|                            |                   | Annual     | days   | 1                     |
| No. of Growing Degree Days | $GDD_{total}$     | Monthly    | days   | 12                    |
|                            |                   | Annual     | days   | 1                     |
| <b>Total</b>               | -                 | -          | -      | <b>117</b>            |

**Dataset partitioning:**

After filtering out the incomplete and missing data points and extracting 117 agro-climatic features, the collective dataset was partitioned into several categories based on geographical information and irrigation water management practices, as summarized in Table 2.

**Table 2. Summary of Dataset Partitioning by Agricultural District and Water Management Practices in Kansas**

| Dataset Type           | District(s) Included | Irrigation Water Source    | Number of Data Points |
|------------------------|----------------------|----------------------------|-----------------------|
| All-Districts Combined | Districts 10-90      | Both Irrigated and Rainfed | 3353                  |
| District-Specific      | District 10          | Irrigated                  | 252                   |
|                        | District 20          | Irrigated                  | 252                   |
|                        | District 30          | Irrigated                  | 444                   |
|                        | District 40          | Irrigated                  | 356                   |
|                        | District 50          | Irrigated                  | 350                   |
|                        | District 60          | Irrigated                  | 431                   |
|                        | District 70          | Rainfed                    | 347                   |
|                        | District 80          | Rainfed                    | 465                   |
|                        | District 90          | Rainfed                    | 456                   |
| Irrigation Districts   | Districts 10-60      | Irrigated                  | 2085                  |
| Rainfed Districts      | Districts 70-90      | Rainfed                    | 1268                  |

After the datasets were created, the mRMR feature selection algorithm was then applied separately to the all-districts combined dataset, individual district-specific datasets, and both irrigation and rainfed district groups to rank all the agro-climatic features as per their relevance to alfalfa yield under different environmental conditions.

### 2.3 Feature Selection Process

Feature selection in this study was applied to identify a minimal subset of agro-climatic predictors that maximized the mutual information shared with observed alfalfa yield variability across spatial domains<sup>52,57</sup>. A total of 117 agro-climatic predictors were included, spanning annual and monthly metrics of temperature and moisture conditions. To control redundancy and retain predictors contributing meaningful information to yield variability, feature selection was implemented in two stages. First, predictors were ranked according to their relevance to yield and their redundancy with other features, assessed using correlation-based statistics derived from the combined datasets. Second, ranked features were evaluated using regression models within a forward selection framework, retaining those that improved predictive performance<sup>58</sup>.

Feature selection methods can generally be categorized into three main types: filter, wrapper, and embedded methods<sup>59,60</sup>. mRMR, a filter-based feature selection method, was used where each predictor relevance to observed yield variability was quantified using mutual-information measures without involving a learning algorithm<sup>51,61</sup>. Wrapper and<sup>52,60</sup> embedded methods perform feature selection during the model training process, as seen in

decision trees or Lasso regression<sup>59</sup>. Because mRMR is a filter-based, model-independent method, it was selected to decouple feature ranking from downstream regression models. This enables a single, consistent feature ranking to be reused across different prediction pipelines, improving reproducibility and flexibility. In addition, its reliance on pairwise relevance and redundancy measures makes it computationally scalable for high-dimensional feature sets. Feature relevance scores reflect shared information with observed yield variability and do not imply effect size or causality. RMSE was calculated for feature subset to quantify model evaluation performance and was used to guide the selection of the subset that provided the most stable representation of observed yield variability.

### 2.3.1 Correlation Analysis of Predictors

To examine potential relationships between predictors and yield, both Pearson (linear) and Spearman (monotonic) correlations<sup>52,62</sup> of all predictor variables with respect to yield were calculated (see details in the supplementary materials section). The corresponding correlation coefficients were in the range of -0.45 to 0.34 (Pearson) and -0.44 to 0.34 (Spearman), with 'tdmean\_Jun\_Total' (total average dew point temperature in June) being the most negatively correlated (-0.45) variable and 'vpdmax\_May\_Total' (total maximum vapor pressure deficit in May) being the most positively correlated (0.34) variable (Fig. S1). The Pearson and Spearman coefficients were highly consistent (correlation = 0.9913), and the top 10 features selected based on both measures were identical.

The overall low correlation values indicate that most predictors exhibit weak or no monotonic relationship with yield, suggesting that simple linear models may be insufficient to capture the underlying relationships<sup>63</sup>. As a result, feature selection methods designed for linear models, such as LASSO, may not be appropriate for this dataset<sup>51,52</sup>. Furthermore, model-dependent approaches may restrict downstream modeling flexibility. Therefore, selecting a filter method (such as mRMR) provides the flexibility of incorporating it with any type of yield prediction and analysis model (both the deterministic and indeterministic ones)<sup>57</sup>.

To leverage this flexibility and capture potential non-linear relationships, we employed three ML-based regression algorithms (see more details in section 2.3.4). These models are capable of learning complex patterns and interactions in the data without assuming linearity, which is particularly suitable given the weak correlations observed.

### 2.3.2 Minimum Redundancy Maximum Relevance (mRMR) algorithm

The mRMR analysis was performed on the dataset to identify the most informative subset of agro-climatic predictors for alfalfa yield. This data-driven, model-free, filter-type algorithm selects features based on two criteria: (i) their individual relevance to the target variable (alfalfa yield) and (ii) their redundancy with other selected features. In this study, these criteria are measured using pairwise mutual information (MI) between each feature and the target yield, and among the features themselves.

The MI between any two continuous variables (e.g., an agro-climatic predictor and yield) quantifies their statistical interdependency and can be expressed as follows.

$$I(u,v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(u,v) \log \frac{p(u,v)}{p(u)p(v)} du dv \quad (4)$$

where,  $p(u)$  and  $p(v)$  are the probability distribution of the variables  $u$  and  $v$ , whereas  $p(u,v)$  is their joint probability distribution, respectively.

In this study, two categories of  $u$  and  $v$  were considered: (i)  $u$  is the target alfalfa yield and  $v$  all 117 agro-climatic variables and (ii) both  $u$  and  $v$  are pairwise combinations of all 117 agro-climatic variables.

A higher value of  $I(u,v)$  suggests stronger dependency, whereas, as the dependence between  $u$  and  $v$  decreases, the mutual information  $I(u,v)$  tends to zero more<sup>64</sup>. For example, the MI between 'tdmean\_Aug\_Total' (the total dew point temperature in August) and the alfalfa yield was calculated as 0.1355, which was much higher than the mean MI (0.0521) between yield and all variables. This indicates that 'tdmean\_Aug\_Total' has a high relevance with yield. On the other hand, 'vpdmin\_Jan\_Total' (the total minimum vapor pressure deficit in January) showed an MI of only 0.0088, indicating that 'vpdmin\_Jan\_Total' might not be strongly associated with yield variation, which aligns with limited crop growth in January. Intuitively, if  $u$  and  $v$  are completely independent,  $I(u,v) = 0$ .

Using this concept of mutual information, mRMR aims to find the best set of  $m \geq 1$  predictors (in descending order) that jointly have the largest dependence on the target  $Y$  as follows.

$$S \ni X^i, i = 1, \dots, m \quad \max D(S,Y) \quad (5)$$

where,  $D(S,Y)$  is the dependency of  $Y$  on set of features,  $S \ni X^i, |S| = m$ . The dependency itself is approximated as the mean value of the mutual information,  $I(X^i,Y)$ , between  $Y$  and each individual member of  $S$  as follows.

$$D(S,Y) = \frac{1}{m} \sum_{X^i \in S} I(X^i,Y) \quad (6)$$

For example, if  $S$  a set of only 2 features,  $S = \{\text{ppt\_Apr\_Total}, \text{vpdmax\_May\_Total}\}$ , and  $I(\text{ppt\_Apr\_Total}, Y) = 0.0598$ ,  $I(\text{vpdmax\_May\_Total}, Y) = 0.1108$ , then  $D(S,Y) = 0.0853$ . However, in this study, the mean MI was found to be 0.0521.

At the same time, the same concept of mutual information is used to minimize the size of  $S$  as follows to ensure that the smallest number of most effective features are selected.

$$\min_S R(S) = \frac{1}{m^2} \sum_{X^i, X^j \in S} I(X^i, X^j) \quad (7)$$

Therefore, combining the above two criteria (as shown in eqns. 5 and 7), the mRMR algorithm optimizes the following objective function:

$$\max_S D(S,Y) - R(S) \quad (8)$$

To illustrate how the mRMR algorithm balanced relevance and redundancy, consider two features from the given dataset: 'tdmean\_annual\_Total' (the annual total dew point temperature) and 'tdmean\_Jun\_Total' (the total dew point temperature in June). Both were moderately relevant to alfalfa yield, with MI values of  $I(\text{tdmean\_annual\_Total}, Y) = 0.125$  and  $I(\text{tdmean\_Jun\_Total}, Y) = 0.130$ , respectively. However, the redundancy between these two features was high with MI  $I(\text{tdmean\_annual\_Total}, \text{tdmean\_Jun\_Total}) = 0.593$ .

In such cases, although 'tdmean\_annual\_Total' is informative, mRMR prioritized 'tdmean\_Jun\_Total' (which is slightly more relevant to the yield) and reduced the final score of 'tdmean\_annual\_Total' due to its strong redundancy with the selected feature ('tdmean\_Jun\_Total'). Thus, a key strength of mRMR is that it favors selecting (top-ranking) a variety of predictors such that each of the predictors contributes independent information to the target variable, rather than duplicating information that is already captured.

Finding the optimal  $S$  is NP-hard<sup>65</sup> i.e., it would require trying out exponential combinations of predictors to find the optimal set of them. Therefore, in this study,  $S$  was constructed using forward addition to reduce the computational complexity<sup>31,33,64</sup>.

### 2.3.4 Selection of the final features

The output of the mRMR algorithm for the climate dataset used in this study was a list of all predictors, ranked in order of their relative importance to the target variable. Hence, to select the best  $n$  predictors for each of the 12 datasets (Table 2) separately, we iterated through all the 117 features as per their mRMR ranking (starting from the most important to the least important one) and performed the following steps in each iteration<sup>35,66</sup>.

- **Step 1 (Cross validation):** for  $i$  (such that  $i = 1, 2, \dots, 117$ ) predictors, create a subset,  $S_i$  of the original dataset  $S$  and split it randomly into a 5-fold CV<sup>67</sup>. Thus, for each data partition, four folds were used to train the prediction models (described below), and the 5<sup>th</sup> fold was used for testing. This process was repeated five times. Thus, the 12 datasets, the number of datapoints per fold were—670, 50, 50, 88, 71, 70, 86, 69, 93, 91, 417, and 253 with 10, 20, 30, 40, 50, 60, 70, 80, and 90 of districts, respectively.
- **Step 2 (Evaluation models):** for each set of  $i$  predictors,  $S_i$ , evaluate its relevance to the yield using three widely used ML-based prediction models<sup>56</sup>—support vector machines (SVM)<sup>45,46</sup>, K-nearest neighbors (KNN)<sup>47,48</sup>, and random forest (RF)<sup>49,50,68</sup>, respectively. The RMSE score<sup>52</sup> was used as an evaluation metric for the model prediction performances<sup>56</sup>. Each of these prediction models was trained using four folds, and the test RMSE was calculated for the 5<sup>th</sup> fold. This process was repeated five times, and for each model, the average of these five simulations was considered as its corresponding evaluation score (see Fig. S6 and Table 3 for details). Then, for the set of predictors  $S_i$ , the final score was calculated as an average of the three prediction models (denoted as ‘Average Value’ in the figures).
- **Step 3 (First local minimum of the moving average):** The moving average of the final scores (average value) was calculated and all the local minima were calculated. Then, the first local minimum (occurring with the least number of predictors) was selected as the ‘best RMSE’ and the corresponding  $S_i$  was selected as the set of best features.

Using Brute force method that checks every possible combination of features, a total of  $1.66 \times 10^{35}$  model evaluations are needed for each dataset which is not possible to simulate with a feasible time. In contrast, for the presented forward feature selection method, three models, 117 different sets of features, and 5-fold CV for each dataset were used. And the total number of model evaluations was  $3 \times 117 \times 5 = 1755$ . Hence, the best set of features were reported based on 21,060 simulations for all 12 datasets (Table 2).

### 2.4 Implementation details

The proposed feature selection method was implemented in MATLAB<sup>69</sup>. All outputs, including the datasets, were later converted into a ‘.csv’ file format. The mRMR algorithm was implemented using the ‘fsrmrmr’ function from Statistics and Machine Learning Toolbox<sup>TM 70</sup>.

For selection of the final features, several functions from Statistics and Machine Learning Toolbox<sup>TM 70</sup> were used. For the 5-fold CV, the ‘cvpartition’ function was used (along with ‘training’ and ‘test’ functions). For the SVM and RF models, the functions ‘fitsvm’ (with the ‘KernelFunction’ specified as ‘rbf’ representing Radial Basis Function<sup>71</sup>) and ‘fitrensemble’ (with the ‘method’ specified as ‘bag’, representing Bootstrap Aggregating<sup>72</sup>) were used, respectively. However, for KNN, we implemented ‘kNNNeighborsRegressor’ function<sup>73</sup> with Euclidean distance specified. The predict function was used for all prediction purposes.

The yield data and some agro-climatic features (temperature, precipitation) were displayed using the surface plot operator (surf), and the mRMR results were plotted using the bar graph operator (bar) on MATLAB. For reproducibility of the results, we used the default value of the 'rng' (Mersenne Twister generator with seed 0) function on MATLAB.

For visualization purposes only (e.g., Fig. 3), missing yield data for certain years and districts were filled using the 1-D data linear interpolation operation (default in 'interp1') in MATLAB, to avoid discontinuities in the plots. This interpolation was performed temporally within each district. internal gaps were estimated from the line drawn between the nearest non-missing years, while edge cases (e.g., missing values in 1981 or 2018) were replaced with the immediate adjacent non-missing year. These interpolated values were not used in the analysis, where missing years and counties were excluded.

### **3 RESULTS AND DISCUSSION**

#### **3.1 Kansas agro-climatic data analysis**

##### **3.1.1 Yield analysis of nine agricultural districts in Kansas (1981–2018)**

Fig. 3 shows the average alfalfa yield from 1981 to 2018. The missing yield data analysis reveals notable disparities in alfalfa yield per acre in districts 10, 20, and 30, where yield consistently exceeds 4 tons/acre. District 30 stands out with yields approaching close to 6 tons/acre in 1990 and 1999 and more than 6 tons/acre in 1998, 2000, and 2001. In contrast, rainfed districts (70–90) consistently demonstrate lower yield production, typically ranging between 1.8 to 4.1 tons/acre throughout the studied period.

During drought occurrences, alfalfa production experiences a corresponding decrease<sup>10</sup>. Notably, Kansas faced a severe drought in 2012<sup>74</sup>, contributing to fluctuating yields over the years. The highest recorded yield, slightly exceeding 6.35 tons/acre, occurred in 2001 (district 30), whereas the lowest yield, just under 1.64 tons/acre, was reported in 2011 (district 60). Based on temperature and precipitation, this study reveals that alfalfa dry matter yield in western Kansas surpassed that of eastern Kansas.

The alfalfa yield differs according to the district, and there was a dramatic difference between districts 10–30 and districts 40–90. This observation, despite gaps in the 2008 dataset, challenges the conventional view that precipitation is the dominant driver of alfalfa growth and instead highlights the predominant role of irrigation in districts 10–60<sup>43</sup>. A comparison of the same counties between 2010 (non-drought) and 2012 (severe drought) confirmed that alfalfa yield increased with higher irrigation inputs during both years, with the effect being more pronounced under drought conditions<sup>75,76</sup>.

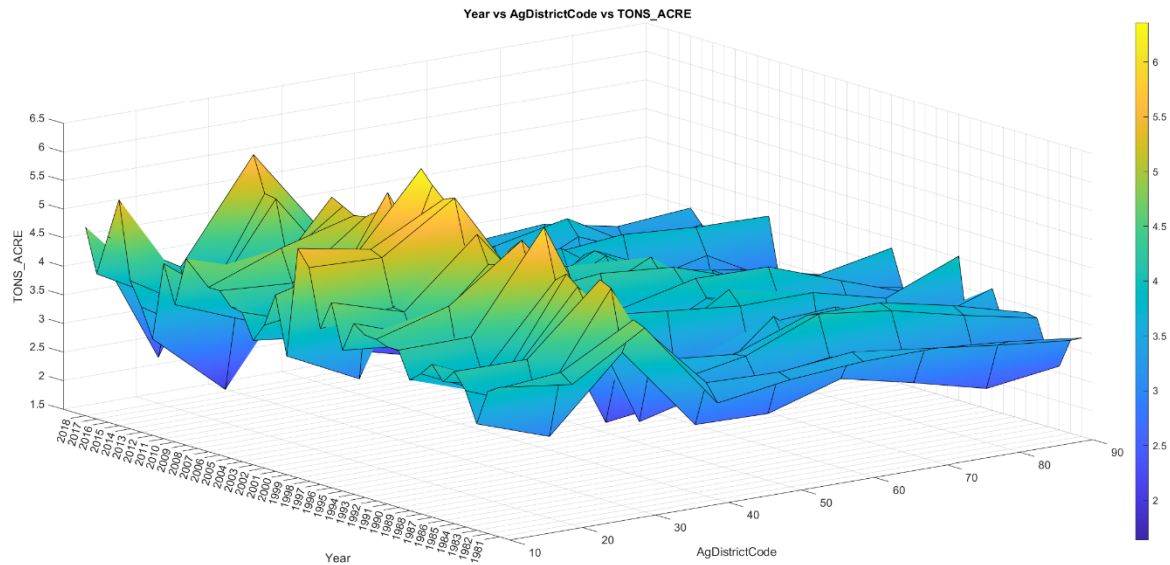


Fig. 3. Average yield for all counties in each district (districts 10–90) for the years 1981–2018, obtained from USDA-NASS. The x-axis represents the years, the y-axis indicates the agricultural districts (10–90), and the z-axis depicts the average yield (tons/acre).

### 3.1.2 Temperature and precipitation analysis of agricultural districts (1981–2018)

For all nine agricultural districts in Kansas (10–90) and years of interest for this study (1981–2018), Fig. 4(a) shows the daily maximum (tmax), average (tmean) and minimum (tmin) temperature. The daily maximum temperature predominantly falls within the range of 95–115 °F across the years (Fig. 4). Conversely, daily average temperature exhibits a variation, mostly between 49–59 °F, with district 90 reaching close to 60.2 °F in 2012 and district 10 dropping to slightly over 48.15 °F in 1993. Furthermore, specific years like 1981, 1986–87, 1990–92, 1998–2001, 2006, and 2011–12 exhibit elevated daily average temperature readings (close to 60 °F), particularly in districts 30, 50, 60, and 90. In contrast, daily minimum temperature displays greater variability (mostly within the range of -30 °F to 5 °F), with the highest being 7.3 °F in district 90 in 2012 and the lowest being -33.6 °F in district 10 in 1989.

Fig. 4(b) presents the daily maximum (pptmax) and average (pptmean) precipitation recorded in inches for districts 10 to 90 and years 1981 to 2018. Daily average precipitation seems to maintain a gradient of precipitation across the districts, with a range of mostly 0.04–0.15 inches. The western districts (10–30) experienced lower precipitation (mostly below 0.07 inches with some exceptions in the years 1993, 1996, 2004–2006, 2009, 2011, 2015, and 2017–2018). Moreover, the year 1993 shows comparatively higher precipitation of at least 0.02 inches than the average for all districts, and an exceptionally high range of precipitation (0.11–0.12 inches) in some of the Central region (districts 40 and 50).

The eastern districts (70–90, also the rainfed districts) show a comparatively higher range of precipitation, mostly above 0.1 inches. For these districts, some of the years (such as 1985–1987, 1992–1993, 1997–1999, 2007–2009, and some areas after 2013) show a higher precipitation amount, mostly in the range of 0.13 to 0.15 inches. In contrast, certain years, such as 1988, 1991, 1994, 2000, 2002, and 2012, experienced a lower range of precipitation, typically around 0.035 to 0.05 inches. However, the year 2006 presented an intriguing scenario where precipitation seemed uniformly distributed across all districts,

exhibiting a narrow variation within the range of 0.062 to 0.083 inches. Remarkably, this resulted in comparatively higher precipitation for the typically low precipitation western districts and lower precipitation for the high precipitation eastern districts during that year.

Moreover, it is important to highlight the existence of several days without any reported precipitation, identified as no precipitation days (nopptdays) and accounted for as a distinct parameter in this study. Overall, considering all days in a year, including those with and without precipitation, the irrigation districts (10–60) averaged daily precipitation of 0.07 inches, whereas the rainfed districts (70–90) averaged 0.11 inches, which is 0.04 inches higher than the irrigation districts.

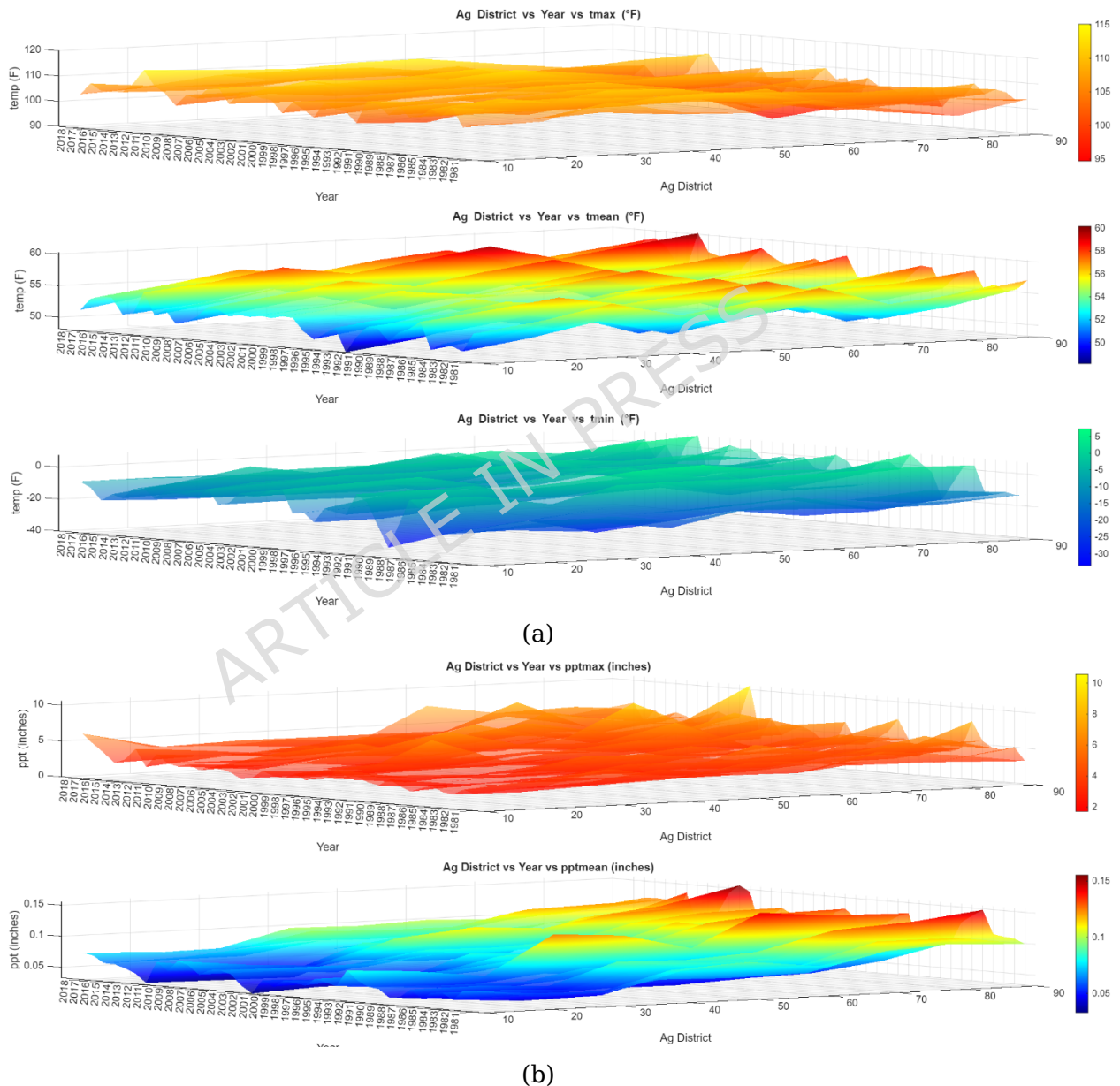
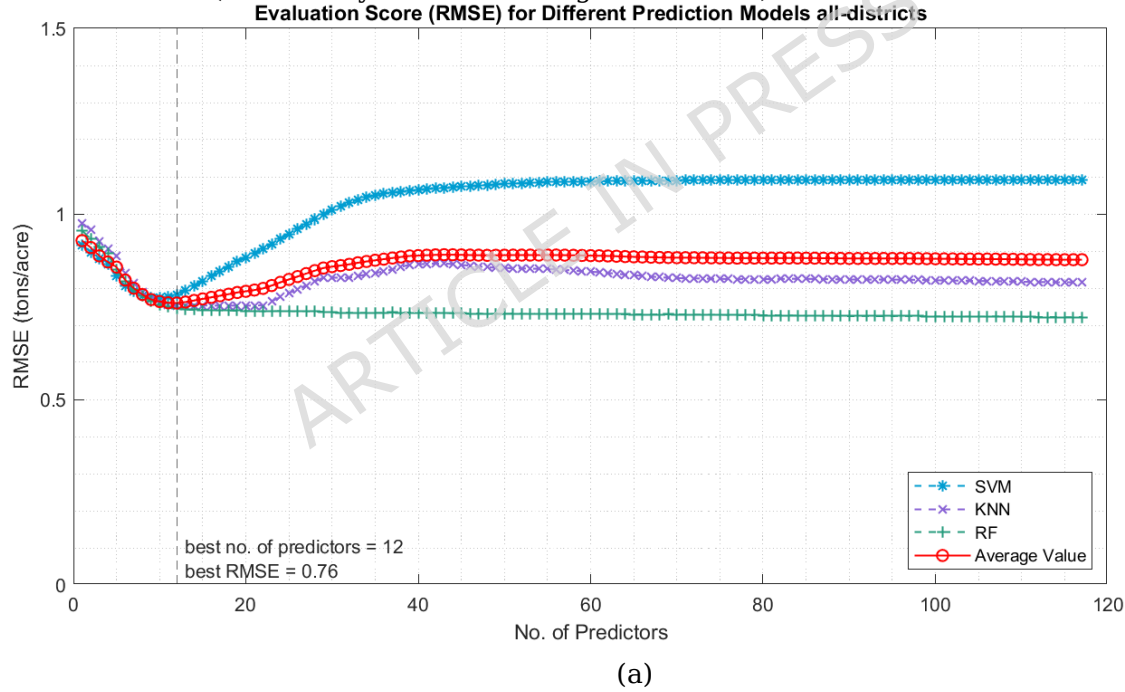


Fig. 4. Maximum (top), mean (middle), and minimum (bottom) daily temperature (a), and maximum (top) and mean (bottom) daily precipitation (b), across counties in districts 10–90 for the period 1981–2018.

### 3.2 Model-based feature selection performance evaluation

The 117 agro-climatic features used in this study (described in the dataset description section) were ranked in order of their importance by the mRMR algorithm individually for each of the 12 datasets (Table 2). Next, each of the feature subsets (obtained in a forward feature selection manner) was evaluated using three ML algorithms—SVM, KNN, and RF, using the default model parameters. These include an RBF kernel with standardized predictors for SVM,  $k = 4$  with Euclidean distance and distance-weighted averaging for KNN, bootstrap aggregation with 100 decision trees and default tree-growing parameters for RF. No hyperparameter tuning was done, as the goal was feature relevance, not predictive optimization<sup>77</sup>. The same forward selection process was repeated for the features ranked in order of their importance by the Pearson's correlation for the Kansas all-districts dataset, which served as a baseline approach.

The performance evaluation results for all-districts datasets are shown in Fig. 5 for (a) mRMR-ranking and (b) Pearson-ranking. Here, the x-axis represents the number of predictors forming the evaluated feature subset. For example, 'No. of Predictors = 1' denotes the highest-ranked feature by mRMR (Fig. 5(a)) and Pearson (Fig. 5(b)) for that dataset. Similarly, 'No. of Predictors = 2' denotes the top two highest-scoring features, and so forth, up to 'No. of Predictors = 117', which includes all available features for the dataset. The y-axis depicts the corresponding test RMSE of the 5<sup>th</sup> fold (as described in the 'selection of the final features' subsection) for the three ML models, along with the average of their RMSE (denoted by the red 'Average Value' line).



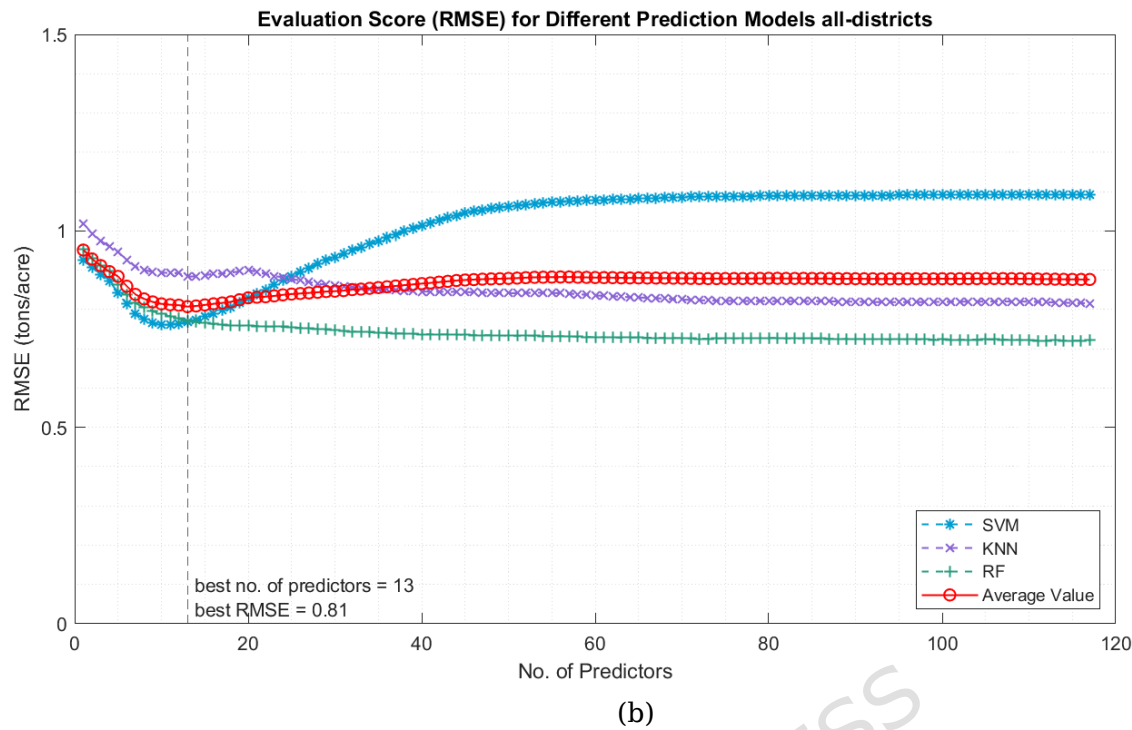


Fig. 5. Average RMSE (tons/acre) for all three ML-based evaluation models (SVM, KNN, RF) on Kansas all-districts dataset using (a) mRMR-ranked features and (b) Pearson-ranked features.

As the mRMR-ranked subfigure shows, with increasing number of features, the evaluation performance used to guide feature subset selection increases (leading to a lower average RMSE) up until 12 predictors with an average RMSE of 0.76 tons/acre by SVM, KNN, and RF. Beyond this point, with an increasing number of predictors, the RF model performance continues to improve marginally by less than 0.01 tons/acre in total, reaching a plateau. But at the same time, SVM performance starts deteriorating in between 8-10 features up to 35-37 features before reaching a plateau at 1.05 tons/acre RMSE at first and then slowly increasing up to 1.09 tons/acre. A similar behavior is observed in the case of KNN, whose RMSE goes as low as 0.75 tons/acre in between 12-22 features and then slowly increasing up to 0.86 tons/acre around 40 features, followed by decreasing and stabilizing at 0.83 tons/acre at around 70 features.

From the Pearson-ranked subfigure, the progression of RMSE across increasing numbers of predictors followed a broadly similar trend to that of the mRMR-ranked features. The optimal average RMSE was obtained with the top 13 Pearson-ranked predictors, yielding 0.81 tons/acre, which is 0.05 tons/acre higher than the mRMR-ranked optimal subset (12 features, 0.76 tons/acre). Model-specific behaviors revealed further contrasts between the two ranking strategies. Under Pearson's ranking, KNN exhibited consistently higher RMSE across nearly all subset sizes compared to the mRMR case, which substantially contributed to the higher overall error. While SVM and RF performances were more comparable between the two methods, both showed slightly elevated RMSE under Pearson ranking. Notably, RF only surpassed the mRMR performance when more than 58 predictors were included, which is far beyond the mRMR optimum of 12 features and Pearson optimum of 13 features.

This decline in prediction accuracy with increasing features illustrates the curse of dimensionality, where sparse high-dimensional data limits model generalization. This effect was evident under both mRMR and Pearson rankings. SVM and KNN, being distance-based models, struggled beyond 10-15 features due to insufficient sample-to-feature ratios. In

contrast, RF remained stable, likely due to its ensemble learning approach. Therefore, as the combined performance of the three ML models reaches its first local minimum at 12 (stops improving before they start deteriorating), the final set of features chosen from the all-districts dataset is the top-scored 12 features by mRMR, as shown in Table 3. From the table, using the proposed approach, the number of selected features varies from 7 (district 80) to 21 (district 10) for different datasets, with a corresponding RMSE of 0.52 to 1.14 tons/acre, respectively. The difference in the number of selected features suggests that district 10 faces a complex environment in which irrigation management interacts closely with localized weather extremes, requiring more features to achieve lower RMSE. In contrast, district 80 is a rainfed system, where alfalfa yield is dominated by fewer climatic drivers, making additional features redundant. Essentially, while more data points in district 80 increase statistical confidence, the underlying simplicity of rainfed production allows the model to capture the variance with only 7 variables.

**Table 3: Number of selected features (as per their mRMR ranking) and the corresponding RMSE score for all 12 datasets for alfalfa production in Kansas.**

| Dataset                      | Number of selected features | RMSE (tons/acre) |
|------------------------------|-----------------------------|------------------|
| All-districts                | 12                          | 0.76             |
| District 10                  | 21                          | 0.71             |
| District 20                  | 9                           | 0.92             |
| District 30                  | 15                          | 1.14             |
| District 40                  | 12                          | 0.52             |
| District 50                  | 11                          | 0.60             |
| District 60                  | 11                          | 0.72             |
| District 70                  | 9                           | 0.63             |
| District 80                  | 7                           | 0.61             |
| District 90                  | 10                          | 0.59             |
| Irrigation districts (10–60) | 10                          | 0.86             |
| Rainfed districts (70–90)    | 10                          | 0.62             |

### 3.3 mRMR selected features

During the summer months, particularly in late July to August, alfalfa growth may be depressed due to the occurrence of higher temperatures in the region. Especially, summer slump is associated with high temperatures, compounded by increased humidity and elevated night temperatures during the summer monsoon<sup>78,79</sup>. Consequently, an increase in respiration rates during warm weather led to decreased net photosynthesis and yield<sup>80</sup>.

Precipitation patterns vary across districts, and they increase from district 10 to district 90 (west to east of Kansas). Western Kansas has been experiencing moderate to exceptional drought for decades from April to October, coinciding with the alfalfa growing season<sup>10,44</sup>. While the eastern region receives more suitable precipitation levels, the western region faces challenges due to insufficient precipitation in Kansas. The factors affecting alfalfa production include the amount of sunlight, length of growing season, temperature, and drought intensity, with precipitation distribution during the growing period crucial for growth, regrowth, and overall yield<sup>10</sup>. Despite varying moisture requirements based on climate and soil conditions, optimal alfalfa cultivation conditions typically necessitate 19.7 to 45.3 inches of water, except for districts 10-30. The water demand for alfalfa production varies, with approximately 4 to 5 inches of water required for each ton/acre of hay production<sup>81</sup>. A high no precipitation day correlates with increased sunshine exposure, indicating more days with higher sun intensity. In irrigated conditions where water needs are met, a day without precipitation can facilitate continuous alfalfa growth, potentially boosting yield. Conversely, in rainfed conditions where moisture is limited, precipitation becomes crucial for alfalfa growth, leading to a positive response to precipitation but a negative impact on the number of no precipitation days variable.

### 3.3.1 mRMR selected variables for the Kansas all-districts dataset

Across all districts in Kansas, twelve variables were selected for their effect on alfalfa production, revealing that the mean dew point temperature of August (Tdmean\_Aug) emerged as the most influential factor (Fig. 6(a)). The mean dew point temperature of August (Tdmean\_Aug) showed a negative relationship with alfalfa yield, which could be due to high temperatures and humidity reducing the evapotranspirative cooling effect from the crop resulting in crop stress (Fig. S8(a)). An increased precipitation in July had significant positive correlation with alfalfa production which is understandable due to the high temperatures and water need in the summer (Fig. S8(a)). Considering the variability in climate among western and eastern Kansas, it is required to understand and address the factors influencing alfalfa yield separately by the districts.

### 3.3.2 mRMR selected variables for the Kansas district-specific datasets

The variables with the highest correlation were number of no precipitation days of September (nopptday\_sep), number of no precipitation days of November (nonpptday\_Nov), Growing Degree Days of April (GDD\_Apr), annual precipitation (ppt\_annual), precipitation of July (ppt\_Jul), mean temperature of July (tmean\_Jul), precipitation of July (ppt\_Jul), minimum VPD of August (VPDmin\_Aug), and maximum VPD of July (VPDmax\_Jul) in districts 10, 20, 30, 40, 50, 60, 70, 80, and 90, respectively. This shows that the variables exerting the highest impact varied across districts, with distinct patterns emerging. Districts 10, 20, 40, 50, and 70 showed a prevalence of the precipitation variable. Meanwhile, variables related to growing degree days and vapor pressure deficit were selected for districts 30, 60, 80, and 90. This difference suggests that the cultivation technique of alfalfa varied among the districts. Specifically, districts 10-60 exhibited a focus on temperature and precipitation variables, while districts 70-90 prioritized precipitation and evapotranspiration variables. Given that irrigation is prevalent in districts 10-60, whereas rainfed conditions prevail in districts 70-90, it underscores the need for distinct analyses tailored to these regions.

From 1890 to 2015, the western region of Kansas was observed to be moisture-limited, which gradually increased due to rising temperatures, increased evaporation rates, low precipitation, and more frost-free days. Meanwhile, the eastern region of Kansas recorded a gradual increase in moisture levels during the same period. In districts 10 to 60, irrigation management shows a statistically negative association with annual precipitation and yield variation related to annual precipitation with a correlation of -0.23 ( $p < 0.05$ ) (Fig. S8(b)). The maximum temperature in the starting of the season and ending of the growing season such as March and October recorded a positive correlation on the alfalfa production. It is plausible that in the irrigated districts (10 to 60), total irrigation water applied per acre during low-precipitation years exceeded the actual crop evapotranspirative demand (Fig. S8(b)). This over-application may have reduced root-zone oxygen availability, increased nutrient leaching, or disrupted osmotic balance, all of which could negatively affect alfalfa growth. Consequently, higher yields in years with higher maximum temperature in the March will help with warming up soil temperatures which will help to break the winter dormancy<sup>82</sup>, and ending of the season in October, it will help in setting optimal time for cutting in dry days in October which will affect subsequent year's alfalfa production<sup>13</sup>. On the other hand, the districts from 70 to 90, where rainfed conditions prevail, precipitation positively impacted production as it had significant positive correlation ( $p < 0.05$ ) (Fig. S8(c)). It was found that an increase in temperature negatively affected production, likely due to increased moisture stress resulting from higher VPD and Tdmean as temperature increases.

### 3.3.3 mRMR selected variables for the Kansas irrigation-districts dataset

In irrigated conditions, the dew point temperature of August (Tdmean\_Aug) and minimum vapor pressure deficit of July (VPDmin\_Jul) had a higher negative correlation

than other variables (Figs. 6(b) and S8(b)). However, most of the variables related to temperature were selected. In particular, the influence related to moisture was high. The dew point temperature of August (Tdmean\_Aug) and minimum vapor pressure deficit of July (VPDmin\_Jul) appeared as the most informative variables associated with yield variation, but these are inversely related, as dew point temperature and vapor pressure deficit were indices of humidity and dryness, respectively. So, if the higher August humidity is known to coincide with reduced summer-season yield potential, consistent with the observed negative associations<sup>26</sup>. The minimum vapor pressure deficit of July was the second-most influential variable under the irrigated conditions. In July, Kansas experiences the highest air temperatures, coupled with dry conditions that pose challenges for meeting alfalfa's water requirements. While irrigation offers a solution to mitigate the challenge, prolonged exposure to high temperatures exceeding 100 °F, combined with standing water for more than 24 hours, can lead to burn damage known as scald injury<sup>26</sup>. Therefore, in irrigated conditions, it was found that many variables related to temperature were selected, indicating that moisture stress is alleviated through irrigation practices<sup>83</sup>.

### 3.3.4 mRMR selected variables for the Kansas rainfed-districts dataset

In the rainfed condition, maximum vapor pressure deficit of August (VPDmax\_Aug) emerged as the most influential factor by negatively affecting alfalfa yield, with most selected variables revolving around vapor pressure deficit, number of no precipitation days, and precipitation parameters (Figs. 6(c) and S8(c)). Alfalfa, known for its high moisture requirements, exhibits a strong dependency to humidity and precipitation levels<sup>9,84,85</sup>. Thus, making maximum vapor pressure deficit of August (VPDmax\_Aug), a critical determinant under rainfed conditions, particularly due to the dry conditions prevalent in August. Despite the significance of precipitation, it was found that variables, namely total precipitation of August (ppt\_Aug\_total), and total precipitation of September (ppt\_Sep\_total), had a low positive impact in the rainfed condition. In rainfed conditions, the normal precipitation levels recorded (1991–2022) were over 29.0 inches<sup>10</sup>, thus satisfying the water requirements for alfalfa growth.

### 3.3.5 mRMR selected key features in different agro-climatic regions in Kansas

In irrigation conditions, the selected variables were in the order - temperature, precipitation, vapor pressure deficit, and dew point temperature (Fig. 6(b)). However, in rainfed conditions the order of variables is different, namely—vapor pressure deficit, number of precipitation days, temperature, and precipitation (Fig. 6(c)). Overlapping variables between irrigated and rainfed conditions included maximum vapor pressure deficit of May (VPDmax\_May) and precipitation of July (ppt\_Jul) (Figs. 6(b) and 6(c)). Maximum vapor pressure deficit of May had significant positive impact in yield ( $p < 0.05$ ) under irrigated conditions, whereas non-significant negative correlation ( $p > 0.05$ ) with alfalfa yield under rainfed conditions (Figs. S8(b) and S8(c)). The maximum vapor pressure deficit of May (VPDmax\_May) is selected likely due to the timing of the first alfalfa harvest in May, when the crop demands ample moisture post-harvest<sup>22</sup>. Meanwhile, precipitation of July (ppt\_Jul) had significant positive correlation with alfalfa yield in both the conditions because the water demand of alfalfa has increased due to the weather in Kansas, which is characterized by high temperature and low precipitation in July.

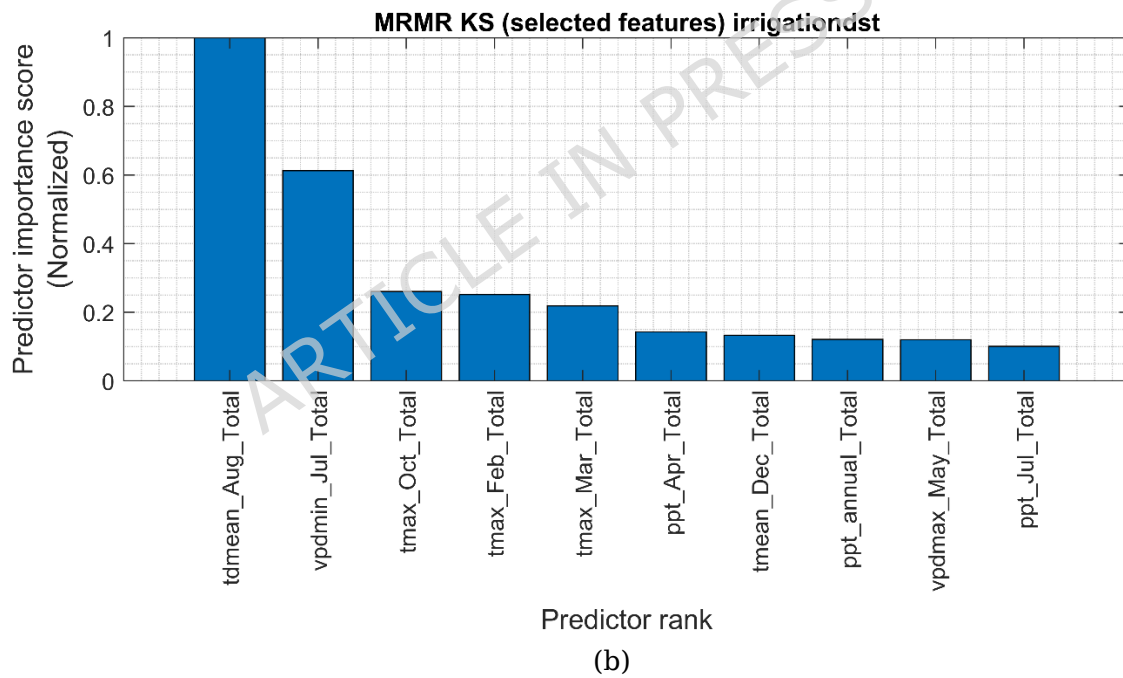
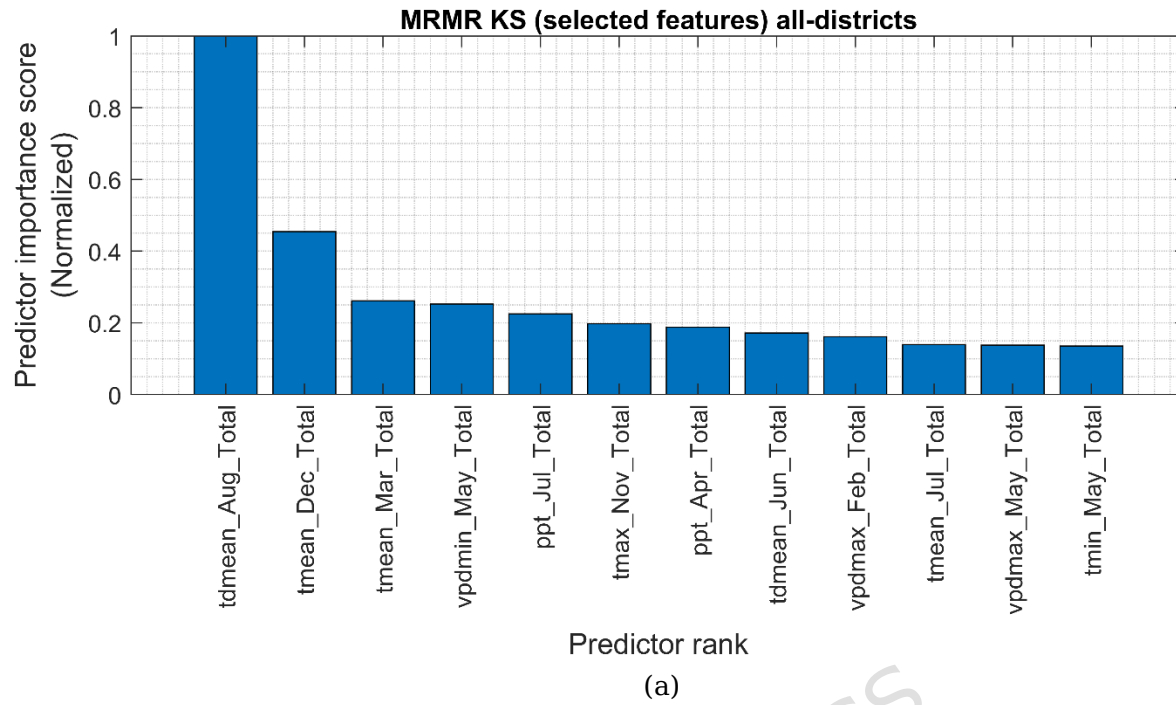
The months selected for both irrigation and rainfed conditions overlapped notably in March, May, July, and August (Figs. 6(b) and 6(c)). The temperature in March was critical for both irrigated and rainfed conditions, highlighting its importance among climatic features. May is crucial as the regrowth of alfalfa begins at this period, and it is dependent on the temperature. On the other hand, in the month of May, alfalfa is harvested, and the vapor pressure deficit variable is considered to be impactful to alfalfa production as it is selected under both rainfed and irrigated conditions because the demand for moisture increases after harvest. July was selected as a precipitation variable under both conditions, which suggests an increased moisture demand for alfalfa due to high temperature. Lastly,

August featured a mean dew point temperature under irrigation, maximum vapor pressure deficit, and precipitation under rainfed, indicating varied factors influencing alfalfa growth during this period.

Several studies have revealed that water is a major yield-limiting factor in alfalfa production<sup>10,86,87,88,89</sup>. While precipitation was initially expected to be the primary influencer in Kansas, it appears to have less impact compared to variables associated with vapor pressure deficit and dew point temperature related to drought and humidity, respectively. This distinction is critical because vapor pressure deficit and dew point act as direct indicators of atmospheric demand and moisture availability, both of which exert continuous influence on crop water status regardless of management. Unlike precipitation, whose effect varies depending on irrigation inputs, VPD and dew point consistently explained yield variability across both irrigated and rainfed systems. Their stability as predictors highlights their value for crop yield modeling, where management-independent drivers are needed. These findings indicate that atmospheric moisture variables consistently carry high informational relevance to observed yield variability, suggesting they warrant further investigation in future alfalfa modeling studies. VPD and dew point are inherently less tied to local management practices. This makes them more transferable predictors across irrigated and rainfed systems, and even across regions with different irrigation technologies or soil types.

The vapor pressure deficit suitable for alfalfa is known to be 3 kPa<sup>90</sup>. Within the timeframe of this study, VPD levels above this threshold appeared intermittently, typically in roughly every alternate year. The same was observed in the Northwest and Midwest regions as the number of 1-day periods above 9 kPa increased by 30–40 days in the late century. It was reported that the dew point temperature significantly increased in the range of 33 to 35 °F in the Midwest region of Kansas<sup>91</sup>. Especially in the Central Midwest, the increase in dew point temperature is highest. Dew point extremes are expected to increase in late summer, adding to climate uncertainty in Kansas<sup>92</sup>.

Under the irrigated conditions, variables primarily linked to humidity and dryness (mean dew point temperature, maximum vapor pressure deficit, and minimum vapor pressure deficit), alongside temperature, were selected, with mean dew point temperature appearing as one of the top-ranked variables by mRMR and exhibiting a negative association with yield based on the correlation analysis. Conversely, rainfed conditions were characterized by variables related to precipitation or dryness (maximum and minimum vapor pressure deficit), with evapotranspiration variables appearing among the most informative predictors in mRMR rankings. Notably, August emerged as the pivotal month influencing both conditions, given its highest impact on alfalfa cultivation due to elevated temperatures and dryness in Kansas.



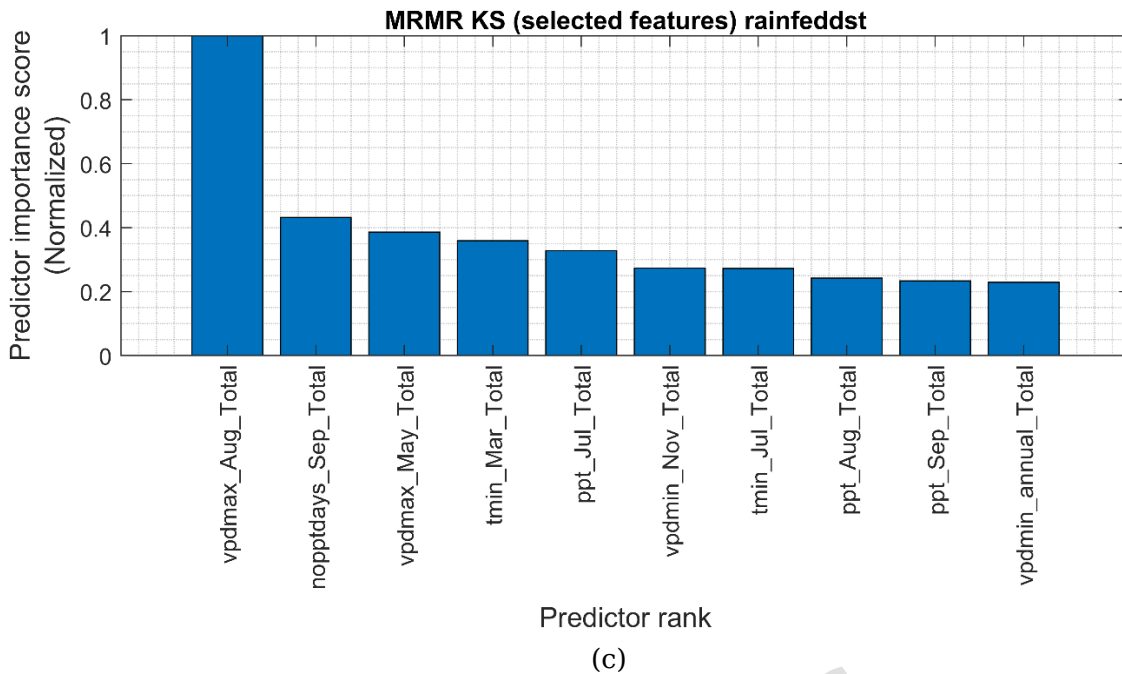


Fig. 6. mRMR selected climate variables affecting alfalfa yield for Kansas all-districts (a), districts under irrigated (b) and rainfed (c) condition in Kansas.

### 3.3.6 mRMR selected top 10 features in comparison to Pearson's and Spearman's correlation coefficients:

For the all-districts datasets, the top 10 features by both Pearson's and Spearman's are: 'tdmean\_Jun\_Total', 'tdmean\_annual\_Total', 'tmin\_Jul\_Total', 'tdmean\_Aug\_Total', 'tdmean\_Apr\_Total', 'tdmean\_Jul\_Total', 'tmin\_Jun\_Total', 'tdmean\_May\_Total', 'tmin\_annual\_Total', 'tmin\_Aug\_Total'. In contrast, the top 10 features by mRMR algorithm represented a more diverse set of predictors, including not only dew point, but also average temperatures, VPD, both maximum and minimum temperatures, and precipitation: 'tdmean\_Aug\_Total', 'tmean\_Dec\_Total', 'tmean\_Mar\_Total', 'vpdmin\_May\_Total', 'ppt\_Jul\_Total', 'tmax\_Nov\_Total', 'ppt\_Apr\_Total', 'tdmean\_Jun\_Total', 'vpdmax\_Feb\_Total', 'tmean\_Jul\_Total', 'vpdmax\_May\_Total', 'tmin\_May\_Total'.

This comparison highlights an important methodological distinction. Correlation-based rankings capture only the strongest marginal associations between individual predictors and yield, which often results in the repeated selection of highly related variables. In contrast, mRMR explicitly balances relevance and redundancy, favoring variables that contribute complementary and non-redundant information to the prediction task. As a result, the feature set identified by mRMR is more heterogeneous and compact, yet ultimately more predictive.

In alfalfa production, it is meaningful to include VPD as a primary variable because it is a direct driver of transpiration and plant water stress, representing the true atmospheric demand for water. This is far more insightful than relying on individual correlations with temperature, which oversimplify the complex plant-environment interaction. An increase in mean dew point in August is statistically correlated with lower yields, but interpreting this correlation requires understanding the underlying physiological cause that is high humidity

(low VPD). When the air is saturated, the VPD is low, which reduces the plant's ability to cool itself through transpiration. The VPD synthesizes the combined effect of temperature and humidity into a single metric that directly reflects the biophysical stress on the plant, providing a more reliable and actionable understanding of yield variability than simple correlation-based rankings.

### 3.3.7 mRMR limitations and future work

While the mRMR algorithm effectively identified climate variables most relevant to alfalfa yield while minimizing redundancy, several limitations should be considered. First, the algorithm relies on pairwise MI, which can be sensitive to data quality, sample size, and variable discretization. District-specific datasets with fewer observations may lead to less stable feature rankings. Second, complex interactions or nonlinear dependencies between variables were only accounted for through MI, not explicitly modeled during feature selection. Future work could explore whether model-specific tuning improves selection sensitivity. Future work may explore other feature selection methods and broader geographic applications. Another limitation of this study is the lack of genetic and genotype-by-environment effects; GDD serves only as a general proxy and does not capture cultivar-specific phenology or stress responses due to dataset constraints. Alfalfa yield is influenced by traits such as dormancy, regrowth, and stress tolerance, which are not represented in the current framework. Future work should couple this ML approach with process-based models like DSSAT-Alfalfa to explicitly integrate genetic responses with climate drivers.

## 4 CONCLUSIONS

This study identified climatic drivers of alfalfa yield using mRMR and ML-based regression. At first, the mRMR algorithm was used to rank all features from most to least important. Next, three ML predictors (SVM, KNN, and RF) were used to evaluate yield variability using 5-fold CV for each dataset with an incremental number of features. Finally, the best set of features was selected based on their corresponding average test RMSE.

To account for climatic and irrigation management differences across the state, the analysis was conducted separately for irrigated and rainfed conditions, in addition to a district-specific as well as a statewide combined dataset. In this regard, variables related to August moisture stress, including mean dew point temperature ( $T_{dmean\_Aug}$ ) and maximum vapor pressure deficit ( $VPD_{max\_Aug}$ ), were ranked as top predictors across datasets. This highlights the relevance of late-summer conditions with alfalfa yield in Kansas.

Under irrigated conditions, the selected variables were temperature, precipitation, vapor pressure deficit, and dew point temperature. However, in rainfed conditions, the selected variables were vapor pressure deficit, precipitation, temperature, and number of no precipitation days. Across all datasets, climatic variables, including VPD and precipitation of certain months, especially March (associated with the onset of vegetative growth), May (around the first harvest), and July–August (peak summer stress period), were identified as the key predictors, emphasizing their importance for alfalfa production.

This data-driven framework helps uncover yield drivers and can be adapted to other crops and regions. Our data-driven approach highlights the importance of management-independent factors like VPD and dew point and offer a useful basis for future climate–yield modeling studies to sustain alfalfa production, providing a transferable framework for resilient crop modeling in drought-prone regions. Future work could extend this framework by integrating remote sensing-derived variables, soil properties, and process-based crop

models to further improve mechanistic understanding and scalability across diverse agro-ecological systems.

### ACKNOWLEDGMENTS

This research was supported by Kansas State University Research and Extension through access to data infrastructure and collaborative resources. Additional support was provided by start-up funding from the College of Agriculture and Life Sciences at North Carolina State University. The authors also acknowledge Montana State University for its contributions towards this work, particularly in developing methodological frameworks and research training. The authors acknowledge Sindhu Palla for partial assistance in outlining the structure of the introduction paragraph related to water issues in the Kansas High Plains.

### AUTHOR CONTRIBUTIONS

F.N. and J.K. contributed to conceptualization, methodology, data curation, formal analysis, and visualization. F.N., J.K., and S.D. wrote the original draft of the manuscript. S.D. also contributed to investigation and formal analysis. B.W. contributed to data visualization, interpretation, and review of the manuscript. D.S. provided statistical guidance, resources, and manuscript editing. D.M. provided agronomic insights and contributed to the review and editing process. G.J. led the conceptualization, supervision, funding acquisition, and project administration, and contributed to manuscript review and editing. All authors reviewed and approved the final manuscript.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### DATA AVAILABILITY

The datasets analyzed in this study are publicly available from the following sources:

Alfalfa yield data: USDA National Agricultural Statistics Service (NASS)

Climate data: PRISM Climate Group (<https://prism.oregonstate.edu/>)

Additional regional precipitation and temperature data: Kansas Mesonet (<https://mesonet.k-state.edu/>)

All data and analysis scripts used in this study are available at the GitHub repository linked below: <https://github.com/BMW-lab-MSU/Alfalfa-Yield-ML-Kansas>

### FUNDING

This study was partially supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under the Alfalfa and Forage Research Program (Grant No. 2023-70005-41080; Accession No. 1031454), as part of the project titled Drought Resilient Alfalfa Production (D-RAP) Using Digital Agriculture and Machine Learning Techniques and Research Capacity Fund (HATCH, HATCH MULTISTATE NC1210), Award No.: 7009808, 7010251..

### REFERENCES

1. Undersander, D. *et al. Alfalfa Management Guide*.

<https://www.agronomy.org/files/publications/alfalfa-management-guide.pdf> (John Wiley & Sons, 2021).

2. Putnam, D. H., Summers, C. G. & Orloff, S. B. Alfalfa production systems in California. in *Irrigated Alfalfa Management for Mediterranean and Desert Zones* 1-19 (University of California, Division of Agriculture and Natural Resource Publication 8287).  
<https://alfalfa.ucdavis.edu/sites/g/files/dgvnsk12586/files/media/documents/UCAlfalfa8287ProdSystems-reg.pdf> (2007).
3. Adhikari, L. & and Missaoui, A. M. Nodulation response to molybdenum supplementation in alfalfa and its correlation with root and shoot growth in low pH soil. *Journal of Plant Nutrition* **40**, 2290-2302.  
[https://www.researchgate.net/publication/318250842\\_Nodulation\\_Response\\_to\\_Molybdenum\\_Supplementation\\_in\\_Alalfa\\_and\\_its\\_Correlation\\_with\\_Root\\_and\\_Shoot\\_Growth\\_in\\_Low\\_Ph\\_Soil](https://www.researchgate.net/publication/318250842_Nodulation_Response_to_Molybdenum_Supplementation_in_Alalfa_and_its_Correlation_with_Root_and_Shoot_Growth_in_Low_Ph_Soil) (2017).
4. McDonald, I., Baral, R. & Min, D. Effects of alfalfa and alfalfa-grass mixtures with nitrogen fertilization on dry matter yield and forage nutritive value. *Journal of Animal Science and Technology* **63**, 305-318.  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8071749/> (2021).
5. Dave, N. 2023 U.S. forage statistics.  
<https://www.progressivepublish.com/downloads/2024/general/2023-pf-stats-lowres.pdf> (2024).
6. Elias, E. H. *et al.* Crop Vulnerability to Weather and Climate Risk: Analysis of Interacting Systems and Adaptation Efficacy for Sustainable Crop Production. *Sustainability* **11**, 6619. <https://www.mdpi.com/2071-1050/11/23/6619> (2019).
7. Whittemore, D. O., Butler Jr, J. J. & Wilson, B. B. 2023 Status of the High Plains Aquifer in Kansas. *Technical Series (Kansas Geological Survey)* **25**, 1-22.  
<https://kgs.ku.edu/sites/kgs/files/files/TechSeriespdfs/TS25.pdf> (2023).
8. Onuoha, H. U. *Time Series Analysis of Long-Term Vegetation Trends, Phenology, and Ecosystem Service Valuation for Grasslands in the US Great Plains*. <https://krex.k-state.edu/items/525a8477-f5f7-4ac5-98d7-e9ba24c5a3fb> (Kansas State University, 2022).

9. Shewmaker, G. E., Allen, R. G. & Neibling, W. H. Alfalfa irrigation and drought.  
*University of Idaho, College of Agricultural and Life Sciences: Moscow, ID, USA*  
<https://efotg.sc.egov.usda.gov/references/public/UT/AlfalfaIrrigationFacts2013Final.pdf>  
(2011).
10. Baral, R., Lollato, R. P., Bhandari, K. & Min, D. Yield gap analysis of rainfed alfalfa in the United States. *Frontiers in plant science* **13**, 931403.  
<https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.931403/full>  
(2022).
11. Dey, S., Jha, G. & Min, D. Precision Irrigation Technologies for Water-Wise and Climate Resilient Alfalfa Production. *Crops & Soils* **57**, 4–11.  
<https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/crso.20383> (2024).
12. Holman, J., Lollato, R. P., Diaz, D. & Min, D. Agronomy eUpdate September 6th, 2019 : Issue 764. [https://eupdate.agronomy.ksu.edu/article\\_new/recommendations-for-fall-planting-of-alfalfa-352](https://eupdate.agronomy.ksu.edu/article_new/recommendations-for-fall-planting-of-alfalfa-352) (2019).
13. McDonald, I., Min, D. & Baral, R. Effect of a fall cut on dry matter yield, nutritive value, and stand persistence of alfalfa. *Journal of Animal Science and Technology* **63**, 799–814.  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8367415/> (2021).
14. F. R. Lamm, K. R. Harmoney, A. A. Aboukheira, & S. K. Johnson. Alfalfa Production with Subsurface Drip Irrigation in the Central Great Plains. *Transactions of the ASABE* **55**, 1203–1212. <https://elibrary.asabe.org/abstract.asp?aid=42258> (2012).
15. Kim, J. Y., Han, K. J., Sung, K. I., Kim, B. W. & Kim, M. Assessment of growing condition variables on alfalfa productivity. *Journal of Animal Science and Technology* **65**, 939–950.  
<https://pubmed.ncbi.nlm.nih.gov/37969334/> (2023).
16. Massmann, A., Gentine, P. & Lin, C. When Does Vapor Pressure Deficit Drive or Reduce Evapotranspiration? *J Adv Model Earth Syst* **11**, 3305–3320.  
<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001790> (2019).
17. Kukal, M. S. & Irmak, S. Nocturnal transpiration in field crops: Implications for temporal aggregation and diurnal weighing of vapor pressure deficit. *Agricultural Water*

- Management* **266**, 107578. <https://www-sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0378377422001251> (2022).
18. Ben-Asher, J., Alpert, P. & Ben-Zvi, A. Dew is a major factor affecting vegetation water use efficiency rather than a source of water in the eastern Mediterranean area. *Water Resources Research* **46**.  
<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008WR007484> (2010).
  19. Luis, I. D., Irigoyen, J. J. & Sánchez-Díaz, M. Low vapour pressure deficit reduces the beneficial effect of elevated CO<sub>2</sub> on growth of N<sub>2</sub>-fixing alfalfa plants. *Physiologia Plantarum* **116**, 497–502. <https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1399-3054.2002.1160408.x> (2002).
  20. Urrego-Pereira, Y. F., Martínez-Cob, A., Fernández, V. & Caverro, J. Daytime Sprinkler Irrigation Effects on Net Photosynthesis of Maize and Alfalfa. *Agronomy Journal* **105**, 1515–1528. <https://acsess.onlinelibrary.wiley.com/doi/10.2134/agronj2013.0119> (2013).
  21. Porter, D. *et al.* Sensitivity of grass-and alfalfa-reference evapotranspiration to weather station sensor accuracy. *Applied Engineering in Agriculture* **28**, 543–549.  
<https://elibrary.asabe.org/abstract.asp?aid=42100&t=2&redir=&redirType=> (2012).
  22. Fink, K. P. *et al.* Alfalfa water productivity and yield gaps in the US central Great Plains. *Field Crops Research* **289**, 108728. <https://www-sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0378429022002994> (2022).
  23. Passioura, J. B. & Angus, J. F. Chapter 2 - Improving Productivity of Crops in Water-Limited Environments. in *Advances in Agronomy* (ed. Sparks, D. L.) vol. 106 37–75.  
<https://www-sciencedirect-com.er.lib.k-state.edu/science/chapter/bookseries/pii/S0065211310060025?via%3Dihub> (Academic Press, 2010).
  24. Konings, A. G., Williams, A. P. & Gentine, P. Sensitivity of grassland productivity to aridity controlled by stomatal and xylem regulation. *Nature Geoscience* **10**, 284–288.  
<https://www.nature.com/articles/ngeo2903> (2017).

25. Kunrath, T. R., Lemaire, G., Sadras, V. O. & Gastal, F. Water use efficiency in perennial forage species: Interactions between nitrogen nutrition and water deficit. *Field Crops Research* **222**, 1–11. <https://www.sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0378429017310729> (2018).
26. Ottman, M. & Mostafa, A. Summer slump in alfalfa. <https://repository.arizona.edu/handle/10150/311219> (2014).
27. Chlingaryan, A., Sukkarieh, S. & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture* **151**, 61–69. <https://www.sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0168169917314710> (2018).
28. Sharma, A., Jain, A., Gupta, P. & Chowdary, V. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access* **9**, 4843–4873. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9311735> (2021).
29. Fazli, S. *et al.* Adaptive crop switching for irrigated agriculture in response to climate change in the Western U.S. *Agricultural Water Management* **318**, 109700. <https://www.sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0378377425004147> (2025).
30. Jeong, J. H. *et al.* Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE* **11**, e0156571. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571> (2016).
31. Thayanandeswari, C. S. S. *et al.* A Machine Learning Approach for Crop Yield Prediction Using Weather Condition. in *2024 International Conference on Advancement in Renewable Energy and Intelligent Systems (AREIS)* 1–5. <https://ieeexplore.ieee.org/document/10893606> (2024).
32. Kumar, C., Dhillon, J., Huang, Y. & Reddy, K. Explainable machine learning models for corn yield prediction using UAV multispectral data. *Computers and Electronics in Agriculture* **231**, 109990. <https://www.sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0168169925000961> (2025).

33. Hammond, K. *et al.* Crop Water Productivity: Within-Field Spatial Variation in Irrigated Alfalfa (*Medicago sativa* L.). *AgriEngineering* **7**. <https://www.mdpi.com/2624-7402/7/4/115> (2025).
34. Noushahi, H. A. *et al.* Selecting High Forage-Yielding Alfalfa Populations in a Mediterranean Drought-Prone Environment Using High-Throughput Phenotyping. *Remote Sensing* **17**. <https://www.mdpi.com/2072-4292/17/9/1517> (2025).
35. Rolnick, D. *et al.* Tackling Climate Change with Machine Learning. *ACM Comput. Surv.* **55**, 42:1-42:96. <https://dl.acm.org/doi/10.1145/3485128> (2022).
36. van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* **177**, 105709. <https://www-sciencedirect-com.er.lib.k-state.edu/science/article/pii/S0168169920302301> (2020).
37. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003* 523-528. <https://pubmed.ncbi.nlm.nih.gov/15852500/> (2003).
38. Mundra, P. A. & Rajapakse, J. C. SVM-RFE With MRMR Filter for Gene Selection. *IEEE Transactions on NanoBioscience* **9**, 31-37. <https://pubmed.ncbi.nlm.nih.gov/19884101/> (2010).
39. Zhang, Y., Ding, C. & Li, T. Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics* **9**, S27. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2559892/> (2008).
40. Wang, G., Lauri, F. & Hassani, A. H. E. Feature Selection by mRMR Method for Heart Disease Diagnosis. *IEEE Access* **10**, 100786-100796. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9894419> (2022).
41. Jo, I., Lee, S. & Oh, S. Improved Measures of Redundancy and Relevance for mRMR Feature Selection. *Computers* **8**, 42. <https://www.mdpi.com/2073-431X/8/2/42> (2019).

42. Shroyer, J. P. *et al.* Alfalfa Production Handbook; Kansas State University Agricultural Experiment Station and Cooperative Extension Service: Manhattan, KS, USA.  
[https://bookstore.ksre.ksu.edu/pubs/alfalfa-production-handbook\\_C683.PDF](https://bookstore.ksre.ksu.edu/pubs/alfalfa-production-handbook_C683.PDF) (1998).
43. Golden, B. B. & Featherstone, A. M. Determinants of Kansas Agricultural Land Values.  
<https://doi.org/10.22004/ag.econ.35033> (2003).
44. Hendricks, N. Estimating irrigation water demand with a multinomial logit selectivity model. <https://krex.k-state.edu/items/cbb9a78c-d206-4fde-b543-df6b251bab9c> (Kansas State University, 2007).
45. Pendell, A. Factors impacting Kansas agricultural land values: 1986-2009.  
<https://krex.k-state.edu/items/f403d9ee-709a-459a-9dbd-a768aefbd624> (Kansas State University, 2013).
46. USDA-NASS. USDA - National Agricultural Statistics Service - Data and Statistics. Data and Statistics. United States Department of Agriculture, National Agricultural Statistics Service. [https://www.nass.usda.gov/Data\\_and\\_Statistics/index.php](https://www.nass.usda.gov/Data_and_Statistics/index.php) (1998).
47. PRISM Climate Group. Time Series Datasets. Northwest Alliance for Computational Science & Engineering (NACSE). <https://prism.oregonstate.edu/explorer/> (2013).
48. Daly, C. *et al.* Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology* **28**, 2031–2064. <https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.1688> (2008).
49. KDA. Irrigation water Use in Kansas, Kansas Department of Agriculture. Decision of Water Resources. Lawrence, KS, U.S.A.  
<https://www.agriculture.ks.gov/home/showpublisheddocument/1908/638457633736200000> (2017).
50. Kansas State University Weather Data Library. Drought Conditions - | Office of the Kansas State Climatologist. <https://climate.ksu.edu/climate/drought/> (2024).
51. Awad, M. & Khanna, R. Support Vector Regression. in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (eds. Awad,

- M. & Khanna, R.) 67–80. [https://link.springer.com/chapter/10.1007/978-1-4302-5990-9\\_4](https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_4) (Apress, Berkeley, CA, 2015).
52. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297. <https://link.springer.com/article/10.1007/BF00994018> (1995).
53. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**, 21–27. <https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf> (1967).
54. Laloë, T. A  $k$ -nearest neighbor approach for functional regression. *Statistics & Probability Letters* **78**, 1189–1193. <https://hal.science/hal-01292692v1/document> (2008).
55. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1. <https://ieeexplore.ieee.org/document/598994> (1995).
56. Prasad, A. M., Iverson, L. R. & Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **9**, 181–199. <https://link.springer.com/article/10.1007/s10021-005-0054-1> (2006).
57. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. & Data, M. Practical machine learning tools and techniques. in *Data mining* vol. 2 403–413. [https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten\\_and\\_Frank\\_DataMining\\_Weka\\_2nd\\_Ed\\_2005.pdf](https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf) (Elsevier Amsterdam, The Netherlands, 2005).
58. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning*. vol. 103. <https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf> (Springer, New York, NY, 2013).
59. Wardlow, B. D., Egbert, S. L. & Kastens, J. H. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains. *Remote Sensing of Environment* **108**, 290–310. <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1001&context=droughtfacpub> (2007).

60. Kansas Mesonet. Kansas Mesonet · Historical Weather. *Kansas Mesonet*  
<https://mesonet.k-state.edu/weather/historical/> (2024).
61. Klocke, N. L., Currie, R. S. & Holman, J. D. Alfalfa response to irrigation from limited water supplies. *Transactions of the ASABE* **56**, 1759–1768.  
<https://elibrary.asabe.org/abstract.asp?aid=44107&t=2&redir=&redirType=> (2013).
62. Sihi, D., Dari, B., Kuruvila, A. P., Jha, G. & Basu, K. Explainable Machine Learning Approach Quantified the Long-Term (1981–2015) Impact of Climate and Soil Properties on Yields of Major Agricultural Crops Across CONUS. *Front. Sustain. Food Syst.* **6**.  
<https://www.frontiersin.org/journals/sustainable-food-systems/articles/10.3389/fsufs.2022.847892/full> (2022).
63. Liu, H. Feature Selection. in *Encyclopedia of Machine Learning* 402–406.  
[https://link.springer.com/rwe/10.1007/978-0-387-30164-8\\_306](https://link.springer.com/rwe/10.1007/978-0-387-30164-8_306) (Springer, Boston, MA, 2011).
64. Marcinkevičs, R. & Vogt, J. E. Interpretability and Explainability: A Machine Learning Zoo Mini-tour. Preprint at <https://doi.org/10.48550/arXiv.2012.01805> (2023).
65. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **40**, 16–28.  
<https://www.sciencedirect.com/science/article/abs/pii/S0045790613003066> (2014).
66. Jović, A., Brkić, K. & Bogunović, N. A review of feature selection methods with applications. in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 1200–1205.  
[https://www.researchgate.net/publication/308871570\\_A\\_review\\_of\\_feature\\_selection\\_methods\\_with\\_applications](https://www.researchgate.net/publication/308871570_A_review_of_feature_selection_methods_with_applications) (2015).
67. Robnik-Šikonja, M. & Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* **53**, 23–69.  
<https://link.springer.com/article/10.1023/A:1025667309714> (2003).

68. Ali Abd Al-Hameed, K. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications* **13**, 3249–3255. [https://ijnaa.semnan.ac.ir/article\\_6079.html](https://ijnaa.semnan.ac.ir/article_6079.html) (2022).
69. Ooi, L., Ng, K.-L., Heng, W.-C., Chua, Y.-F. & Lim, Y.-H. Human resource practices and employee engagement: A study among academicians in Malaysian private higher education institutions. *International Journal of Management Studies* **29**, 71–100. <https://www.semanticscholar.org/reader/6aab293faf273f46984863d09245db598bdc2e49> (2022).
70. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 1226–1238. <https://pubmed.ncbi.nlm.nih.gov/16119262/> (2005).
71. Thomas H, C., Charles, E., Ronald L, R. & Clifford, S. Introduction to algorithms third edition. <https://www.cs.mcgill.ca/~akroitt/math/compsci/Cormen%20Introduction%20to%20Algorithms.pdf> (2009).
72. Zhao, Z., Anand, R. & Wang, M. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* 442–452. <https://ieeexplore.ieee.org/document/8964172> (2019).
73. Witten, I. H., Frank, E. & Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. <https://tjzhifei.github.io/links/DM3.pdf> (Morgan Kaufmann, Burlington, MA, 2011).
74. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32. <https://link.springer.com/article/10.1023/A:1010933404324> (2001).
75. The MathWorks Inc. The MathWorks Inc. (2023).
76. The MathWorks Inc. Statistics and Machine Learning Toolbox Documentation. <https://www.mathworks.com/help/stats/index.html> (2023).

77. Kecman, V., Huang, T.-M. & Vogt, M. Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance. in *Support Vector Machines: Theory and Applications* (ed. Wang, L.) 255-274.  
[https://www.researchgate.net/publication/227228113\\_Iterative\\_Single\\_Data\\_Algorithm\\_for\\_Training\\_Kernel\\_Machines\\_from\\_Huge\\_Data\\_Sets\\_Theory\\_and\\_Performance](https://www.researchgate.net/publication/227228113_Iterative_Single_Data_Algorithm_for_Training_Kernel_Machines_from_Huge_Data_Sets_Theory_and_Performance) (Springer, Berlin, Heidelberg, 2005).
78. Breiman, L. Bagging predictors. *Mach Learn* **24**, 123-140.  
<https://link.springer.com/article/10.1007/BF00058655> (1996).
79. Ferreira, D. k-Nearest Neighbors (kNN) Regressor (1.0.1). *GitHub*  
<https://github.com/ferreirad08/kNNNeighborsRegressor/releases/tag/1.0.1> (2020).
80. NDMC. United States Drought Monitor. Time Series Data on Kansas Percent Area in U.S. Drought Monitor Categories. The National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration.  
<https://droughtmonitor.unl.edu/DmData/TimeSeries.aspx> (2022).
81. KDA. Kansas Irrigation Water Use 2010, Kansas Department of Agriculture. Decision of Water Resources. Lawrence, KS, U.S.A.  
<https://www.agriculture.ks.gov/home/showpublisheddocument/1896/638457633692300000> (2010).
82. KDA. Kansas Irrigation Water Use 2012, Kansas Department of Agriculture. Decision of Water Resources. Lawrence, KS, U.S.A.  
<https://www.agriculture.ks.gov/home/showpublisheddocument/1902/638457633715430000> (2012).
83. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182.  
<https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf> (2003).

84. Feltner, K. C. & Massengale, M. A. Influence of Temperature and Harvest Management on Growth, Level of Carbohydrates in the Roots, and Survival of Alfalfa (*Medicago sativa* L.). *Crop Science* **5**. cropsoci1965.0011183X000500060030x (1965).
85. Robison, G. D. & Massengale, M. A. Effect of Harvest Management and Temperature on Forage Yield, Root Carbohydrates, Plant Density and Leaf Area Relationships in Alfalfa (*Medicago sativa* L. cultivar 'Moapa'). *Crop Science* **8**. cropsoci1968.0011183X000800020002x (1968).
86. Brown, R. H., Pearce, R. B., Wolf, D. D. & Blaser, R. E. Energy Accumulation and Utilization. in *Agronomy Monographs* (ed. Hanson, C. H.) 143-166.  
<https://access.onlinelibrary.wiley.com/doi/10.2134/agronmonogr15.c7> (American Society of Agronomy, Madison, WI, USA, 2015).
87. Sloan, J. How Much Water Does Alfalfa Need? *Extension / University of Nevada, Reno*  
<https://extension.unr.edu/publication.aspx?PubID=2575> (2009).
88. Lang, B. Plan to Scout Alfalfa Stands for Winter Survival. *Integrated Crop Management*  
<https://crops.extension.iastate.edu/cropnews/2017/03/plan-scout-alfalfa-stands-winter-survival> (2017).
89. Undersander, D. Economic Importance, Practical Limitations to Production, Management, and Breeding Targets of Alfalfa. in *The Alfalfa Genome* (eds. Yu, L.-X. & Kole, C.) 1-11.  
[https://www.researchgate.net/publication/353318504\\_Economic\\_Importance\\_Practical\\_Limitations\\_to\\_Production\\_Management\\_and\\_Breeding\\_Targets\\_of\\_Alalfa](https://www.researchgate.net/publication/353318504_Economic_Importance_Practical_Limitations_to_Production_Management_and_Breeding_Targets_of_Alalfa) (Springer International Publishing, Cham, 2021).
90. Takele, E. & Kallenbach, R. Analysis of the Impact of Alfalfa Forage Production under Summer Water-Limiting Circumstances on Productivity, Agricultural and Growers Returns and Plant Stand. *J Agronomy Crop Science* **187**, 41-46.  
<https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1439-037X.2001.00501.x> (2001).
91. Schneekloth, J. & Andales, A. Seasonal water needs and opportunities for limited irrigation for Colorado crops. *Colorado State University Extension Fact Sheet* **4**.

- <https://extension.colostate.edu/resource/seasonal-water-needs-and-opportunities-for-limited-irrigation-for-colorado-crops/> (2017).
92. Jáuregui, J. M. *et al.* Yield gaps of lucerne (*Medicago sativa* L.) in livestock systems of Argentina. *Annals of Applied Biology* **181**, 22–32.  
<https://onlinelibrary.wiley.com/doi/abs/10.1111/aab.12745> (2022).
93. Holzman, M. E. & Rivas, R. E. Early maize yield forecasting from remotely sensed temperature/vegetation index measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **9**, 507–519.  
<https://ieeexplore.ieee.org/document/7378859?denied=> (2016).
94. Carter, P. R. & Sheaffer, C. C. Alfalfa Response to Soil Water Deficits. I. Growth, Forage Quality, Yield, Water Use, and Water-Use Efficiency. *Crop Science* **23**.  
crops1983.0011183X002300040016x (1983).
95. Kilcher, M. R. & Heinrichs, D. H. Stand patterns for alfalfa-grass hay production in a dry climate. *Can. J. Plant Sci.* **51**, 317–322.  
<https://cdnsiencepub.com/doi/abs/10.4141/cjps71-061> (1971).
96. Gamelin, B. L. *et al.* Projected U.S. drought extremes through the twenty-first century with vapor pressure deficit. *Sci Rep* **12**, 8615. <https://www.nature.com/articles/s41598-022-12516-7> (2022).
97. Scheff, J. & Burroughs, J. C. Diverging trends in US summer dewpoint since 1948. *International Journal of Climatology* **43**, 4183–4195.  
<https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.8081> (2023).
98. Brown, P. J. & DeGaetano, A. T. Trends in U.S. Surface Humidity, 1930–2010.  
<https://doi.org/10.1175/JAMC-D-12-035.1> (2013).