Integrating physiology and machine learning: a novel framework for precision wheat management in water-limited environments

Mohsen Jahan[1], Mohammad Behzad Amiri[2*], Ainollah Hesami[3], Atefeh Mirzaeian[4]

[1] Department of Agrotechnology, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Department of Plant Production, Faculty of Agriculture, University of Gonabad, Gonabad, Iran

(Email: Amiri@gonabad.ac.ir)

[3]Department of Plant Production, Faculty of Agriculture, University of Shoushtar, Shoushtar, Iran

[4] Department of Agrotechnology, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

This research introduces a novel data-driven framework that combines hierarchical climate zoning with explainable machine learning to forecast county-level wheat yield and identify significant climatic factors in Khorasan Razavi, Iran. Based on 20 years of high-resolution climate data (2004–2023), we identified three agroclimatic zones: semi-arid plains, intermontane semi-arid plains, and arid plains/desert. The study utilized Random Forest and XGBoost classifiers to classify yields into Low, Medium, and High categories, demonstrating strong performance (macro F1: 0.72–0.75). A critical SHAP (SHapley Additive exPlanations) analysis indicated that growing season length (GSL) and minimum temperature (Tmin) were the most reliable yield drivers across clusters, accounting for physiological processes such as grain-filling duration and respiratory losses. In arid regions, model predictions were primarily influenced by precipitation-related a variable, which emphasizes moisture scarcity as a primary limitation. In contrast to conventional correlation-based methods, our explainable machine learning framework offers a mechanistic understanding of yield-climate interactions at a high spatial resolution, facilitating practical strategies like cluster-specific sowing windows and precise irrigation scheduling. This approach presents a scalable decision-support instrument for climate adaptation in wheat-producing arid regions Khorasan Razavi, and it supports process-based models by providing both transparency and practicality in data-scarce situations.

Introduction

Globally, agricultural systems are becoming increasingly susceptible to climate variability and extreme weather events like persistent droughts and heatwaves, thereby endangering food security and socio-economic stability (IPCC, 2023). Temporal and spatial variations in yield are influenced by input factors, agricultural methods, soil composition, and climatic factors. Previous agroclimatological studies from the past four decades, which used linear regression or simple correlation analyses to examine the impact of temperature and growing season duration on wheat yield variability, were limited by their spatial resolution and interpretability compared to contemporary machine learning frameworks. This research contributes to the field by combining hierarchical climate clustering with ensemble machine learning models (Random Forest and XGBoost) and SHAP analysis (Nassiri-Mahallati and Jahan, 2020).

The AquaCrop model's reliability in simulating soil water content, sesame canopy cover, and final yield was confirmed by sensitivity analysis (Nassiri-Mahallati and Jahan, 2020), emphasizing the viability of model-based methods in arid environments. This research introduces a scalable and transferable framework for climate-smart wheat management in arid and semi-arid environments, utilizing the integration of clustering with machine learning–based classification.

Materials and methods

Khorasan Razavi Province, situated in northeastern Iran (33°52′–37°42′ N, 56°19′–61°16′ E) and spanning 118,854 km², is characterized by a largely semi-arid climate. This research examines the correlation between agroclimatic factors and wheat production within the province. The study involved the development of a 20-year dataset (2004–2023) that incorporated yearly county-level wheat yield information from official agricultural archives with seventeen critical agroclimatic indices. To ensure full spatial and temporal coverage, data from 19 Iranian Meteorological Organization (IRIMO) stations was supplemented, as required, with information from the closest available station. Where direct observation data was not accessible, or to address small data gaps (less than 5% of the total records), high-resolution reanalysis datasets (AgMERRA, APHRODITE, ERA5) were utilized. The efficacy of these datasets in modeling the region is verified through multiple studies that indicate excellent performance (e.g., temperature R^2 exceeding 0.90). The Aridity Index (AI) was computed as the quotient of annual precipitation and potential evapotranspiration, which produced the principal indices. Khorasan Razavi Province was divided into its 19 administrative divisions. Each county was characterized by the long-term average and standard deviation of five critical climate variables: Mean temperature (Tmean), precipitation (Prec), growing degree days (GDD), evapotranspiration (ET), and temagglomerative clustering was conducted utilizing Ward's method, with Euclidean distance, within the perature seasonality (TS). after z-score standardization, hierarchical R software environment.

Results and discussion

SHapley Additive exPlanations (SHAP) analysis provided mechanistic insight into the key climatic drivers of yield class across all models (Figures 2a & 2b).

This shift underscores moisture availability as the primary To complement the SHAP-based global interpretability analysis, we conducted a local driver assessment using standardized deviations (Z-scores) for each climatic feature. Z-scores were calculated based on deviations from the multi-year, county-level mean. While all variables were assessed, only those with deviations of ±0.5 standard deviations or more were considered statistically meaningful. These are referred to as effective drivers—climatic variables that significantly diverged from historical norms in either direction and contributed to yield anomalies. For example, in the highest-yielding year and county (Figure 4), parameters such as growing season length (GSL), maximum temperature during the growing season (TmaxGS), average temperature during the growing season (TmeanGS). Number of days with temperature above 30 OC (NDO30) and evapotranspiration (ET) showed notable positive deviations.

Conversely, in the lowest-yielding scenario (Figure 5), elevated aridity index (AI) and reduced evapotranspiration (ET) emerged as key negative drivers. This complementary analysis reinforces the reliability of model explanations and provides granular insights into local climate–yield relationships that are actionable for site-specific management.

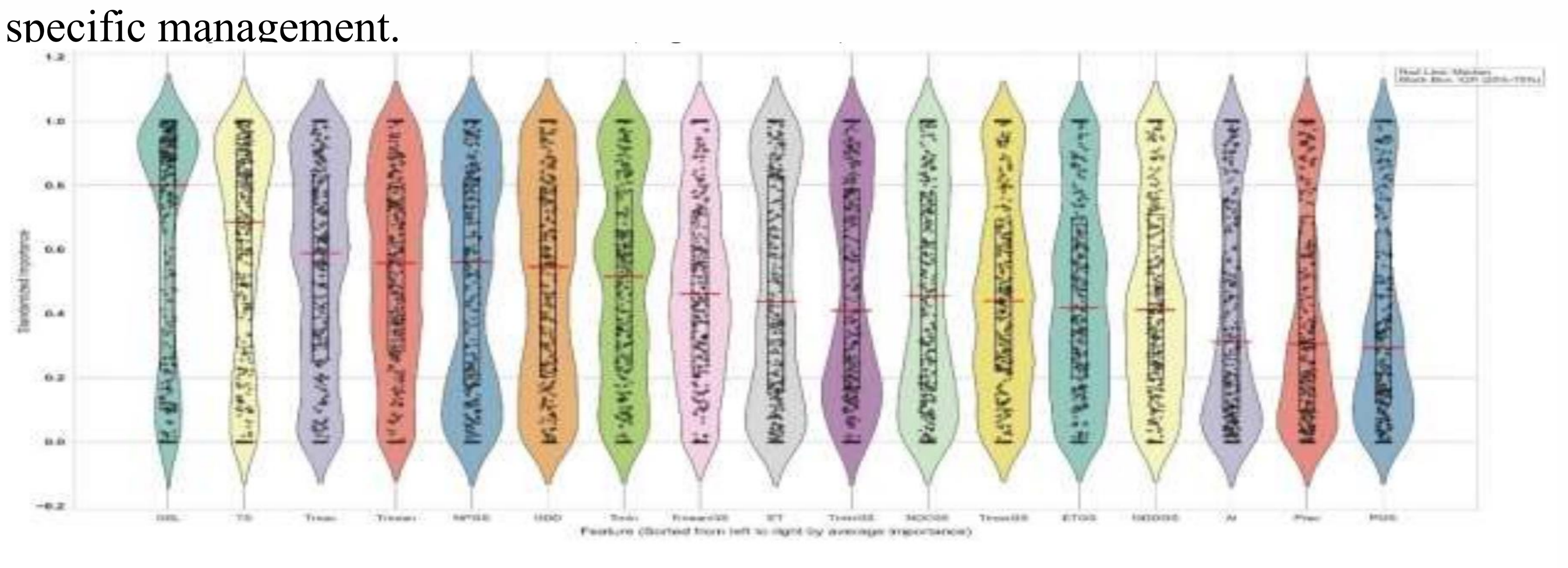

Figure 2. Simplified violin plots of SHAP value distributions for the top five features (GSL, Tmin, NPGS, PGS, TS) in the Random Forest model, illustrating their impact on wheat yield classification in Khorasan Razavi Province.

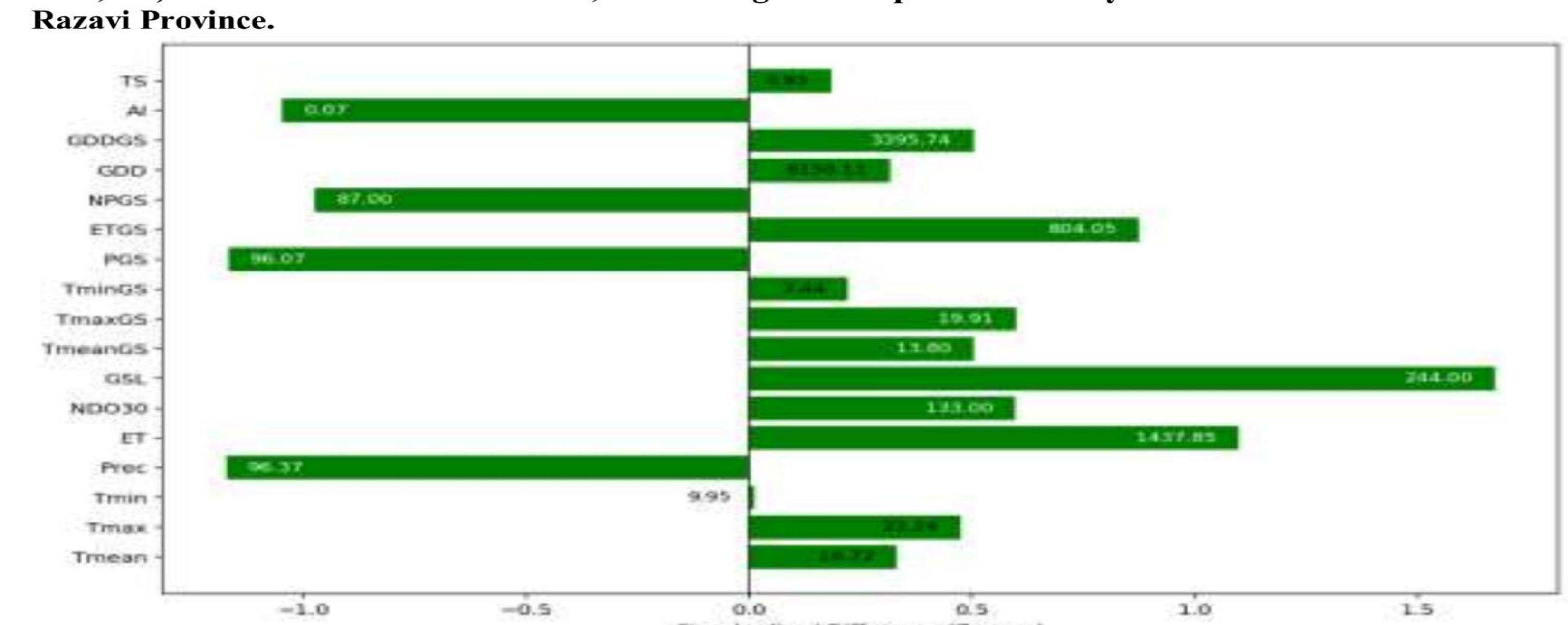

Figure 4. Standardized Climatic Drivers of High Wheat Yield in Torbat-E Heydariyeh (2021)

References

1. Dadrasi, A., Chaichi, M., Nehbandani, A., Soltani, E., Nemati, A., Salmani, F., Heydari, M. and Yousefi, A.R. (2023). Global insight into understanding wheat yield and production through Agro-Ecological Zoning. *Scientific Reports*, *13*(1), 15898. IPCC. (2023). Special Report: Special Report on Climate Change and Land: Ch 05 Food Security.
2. Lesk, C., Coffel, E., Winter, J., Ray, D., Zscheischler, J., Seneviratne, S. I., & Horton, R. (2021). Stronger temperature–moisture couplings exacerbate the impact of climate warming on global crop yields. *Nature food*, *2*(9), 683-691.
3. Nassiri-Mahallati, M., & Jahan, M. (2020). Using the AquaCrop model to simulate sesame performance in response to superabsorbent polymer and humic acid application under limited irrigation conditions. *International Journal of Biometeorology*, *64*(12), 2105-2117.