RESEARCH ARTICLE AGROBIOLOGICAL RECORDS
ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

AGROCLIMATIC VARIABILITY AND MACHINE LEARNING–GIS FORECASTING OF HARMFUL NON-GREGARIOUS LOCUSTS IN KAZAKHSTAN

Kurmet Baibussenov, Arman Rustembayev*, Aigul Bekbayeva, Zhasulan Amanbay, Askar Zhubatkanov and Arsen Dzhumagulov

Saken Seifullin Kazakh Agrotechnical University, Astana, Republic of Kazakhstan

*Corresponding author: rustem.arman88@gmail.com

ABSTRACT

Agroclimatic variability influences harmful non-gregarious locust populations across Kazakhstan, but forecasting performance may differ among ecological zones. This study developed a machine learning–GIS framework using 7,192 phytosanitary monitoring records collected in 25 agroclimatic zones during 2003–2024. We integrated monitoring data with meteorological and remote-sensing predictors, and classified zones as high, moderate, or low risk based on mean absolute population density (P_{abs}). We evaluated Gradient Boosting, Support Vector Regression (SVR), Random Forest, SARIMA, and Linear Regression using a model-development period of 2003–2018 and an independent test period of 2019–2024. In the high-risk category, Gradient Boosting was selected because it achieved the lowest RMSE (2.636 individuals m^{-2}) and the highest R^2 (0.005), although SVR produced the lowest MAE (1.307 individuals m^{-2}). SVR was selected for the moderate-risk category (MAE = 1.135; RMSE = 3.039 individuals m^{-2}) and the low-risk category (MAE = 1.212; RMSE = 1.940 individuals m^{-2}), although R^2 remained slightly negative in both categories. Exploratory screening associated higher population densities with land-surface temperatures of 30–35°C, precipitation of 20–50 mm, normalized soil-moisture values of 0.60–0.75, and SPI values of 0–0.5. Treat these ranges as dataset-specific warning indicators. The framework can help phytosanitary agencies prioritize field surveys and direct verification toward higher-risk zones, while farmers can use zone-level alerts to guide scouting. Because forecast skill was limited and formal prediction intervals were unavailable, the outputs should support, rather than replace, field-based phytosanitary management decisions.

***Keywords*:** Non-gregarious locusts, Agroclimatic variability, Machine learning, GIS, Phytosanitary forecasting, Kazakhstan.

Article History: ABR-26-197 || Received: 28-May-2026 || Revised: 02-Aug-2026 || Accepted: 09-Aug-2026 || Published Online: 11-Sep-2026

1. INTRODUCTION

Locust forecasting supports food-security planning, but the forecasting target differs fundamentally between the desert locust and harmful non-gregarious acridids (Mamo et al., 2025). Desert locust early-warning systems are designed around density-dependent phase transformation, rapidly changing breeding habitats, long-distance migration, and transboundary swarm movement; therefore, they usually predict presence probability, breeding-area suitability, and dispersal from rainfall, vegetation greening, soil moisture, and field-survey data (Sun et al., 2022; Cornejo-Bueno et al., 2023; Retkute et al., 2024; Marescot et al., 2025). In contrast, the non-gregarious species complex considered in the present study consists of resident populations that persist at background densities and periodically form localized outbreak foci within relatively stable agricultural landscapes (Baibussenov et al., 2022, 2024; Baibussenov & Amanbay, 2025). Forecasting these populations therefore requires estimating local population density and threshold exceedance across heterogeneous agroclimatic zones rather than predicting swarm occurrence or migration. These differences affect the response variable, spatial resolution, temporal horizon, and management objective, preventing the direct transfer of desert-locust forecasting models to non-gregarious pest systems.

1.1. Ecological and Agro-Economic Relevance of Non-Gregarious Locusts

Locusts and grasshoppers (Orthoptera: Acrididae) are among the most ecologically influential herbivorous insects in arid and semi-arid agroecosystems. Foundational acridological synthesis established that locust population dynamics are regulated by density-dependent phase transitions, climatic variability and trophic interactions. Subsequent ecological research demonstrated that grasshopper outbreaks alter primary productivity, nutrient cycling, and rangeland stability.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

In addition to the gregarious locust species *Calliptamus italicus* (Linnaeus, 1758), *Dociostaurus maroccanus* (Thunberg, 1815), and *Locusta migratoria* (Linnaeus, 1758) (Azhbenov et al., 2015), a complex of economically significant non-gregarious acridid species constitutes a persistent phytosanitary threat to agricultural production in Kazakhstan (Baibussenov et al., 2022). Their distribution across the country shows pronounced agroclimatic zonation driven by gradients in moisture availability and thermal resources (Baibussenov et al., 2022). In the forest-steppe and northern steppe zones, characterized by comparatively higher precipitation and moderate temperature regimes, the dominant species include *Gomphocerus sibiricus* (Linnaeus, 1767), *Chorthippus albomarginatus* (De Geer, 1773), *Oedaleus decorus* (Germar, 1825), *Pararcyptera microptera* (Fischer von Waldheim, 1833), *Podisma pedestris* (Linnaeus, 1758), and *Stauroderus scalaris* (Fischer von Waldheim, 1846). The abundance of these species is closely associated with the productivity and structural composition of grass–forb communities.

In the dry-steppe and semi-desert regions, where the climate is strongly continental and characterized by periodic droughts, *Dociostaurus brevicollis* (Eversmann, 1848), *Dociostaurus kraussi* (Ingenitsky, 1897), and *Oedipoda caerulescens* (Linnaeus, 1758) predominate. These species show high ecological plasticity and can form localized outbreak foci under favorable spring soil moisture conditions. In the semi-desert and desert zones of southern and western Kazakhstan, Calliptamus turanicus (Tarbinsky, 1930), Calliptamus barbarus (Costa, 1836), Oedipoda miniata (Pallas, 1771), and Ramburiella turcomana (Fischer von Waldheim, 1833) form stable populations. The population density of these taxa is primarily regulated by cumulative heat supply and soil moisture dynamics during embryonic development.

Long-term ecological studies confirmed that population regulation depends on both abiotic forcing and density-dependent mortality (Mallie et al., 2026). Regional analyses in Asian steppe ecosystems revealed spatial clustering of outbreak foci linked to vegetation heterogeneity. Climatic variability plays a central role in outbreak dynamics. Rainfall anomalies and temperature accumulation explain interannual population fluctuations at continental scales. Coupled vegetation–locust feedbacks further reinforce these dynamics (Retkute et al., 2024), while density-dependent phase transitions amplify demographic responses. Geographic variation in recruitment patterns under projected climate change highlights the vulnerability of steppe ecosystems to outbreak intensification (Humphreys et al., 2022).

Climate-driven ecological responses have been widely documented across terrestrial insect systems. Global climate assessments indicate increasing aridity and thermal variability in continental interiors (Mingalev, 2021), directly affecting agricultural productivity and food security. In Kazakhstan, such climatic instability increases the probability of locust population surges. Kazakhstan provides a particularly informative model region for climate-sensitive locust forecasting because its territory forms a continuous continental gradient from forest-steppe and steppe to dry-steppe, semi-desert, and desert environments. Within a single national monitoring system, these landscapes encompass 25 agroclimatic zones with sharply contrasting thermal and moisture regimes and support different assemblages of harmful non-gregarious acridids. The availability of nationwide phytosanitary observations for 2003–2024 makes it possible to compare model behavior under arid, transitional, and humid conditions using a common population-density indicator. Kazakhstan therefore represents a natural test bed for evaluating the ecological transferability of ML–GIS forecasting across the heterogeneous semi-arid agroecosystems' characteristic of Central Asia (Mingalev, 2021; Baibussenov et al., 2022; Wu et al., 2022; Baibussenov et al., 2024).

1.2. Classical Phytosanitary Forecasting: Structure and Limitations

Traditional phytosanitary monitoring systems rely on standardized accounting methods, field surveys, and expert threshold assessments (Gahramanov et al., 2024). Operational manuals emphasize descriptive extrapolation of previous-year densities and climatic analogues. However, statistical ecology has demonstrated that multicollinearity among environmental predictors biases parametric estimation (Elith and Leathwick 2009). Remote sensing integration into vegetation science further revealed that ecological drivers operate across multiple spatial scales. Deep neural architectures expanded predictive capability in nonlinear systems. Ecological classification tasks confirmed improved robustness of Random Forest in species–environment relationships. In agricultural pest modeling, machine learning significantly reduced forecasting error compared to classical biometeorological models (de Oliveira Aparecido et al., 2020; Singh et al., 2024). Temporal dependencies further necessitate recurrent neural architectures. LSTM networks effectively capture lagged climatic effects (Xiao et al., 2019). Advances in deep learning demonstrate transformative capacity in complex environmental prediction (Reichstein et al., 2019).

1.3. Machine Learning and Spatiotemporal Modeling in Pest Forecasting

Reichstein et al. (2019) emphasized that deep neural networks enable extraction of latent ecological relationships from large, heterogeneous datasets, a key requirement in multi-source phytosanitary forecasting. In the specific context of locust ecology, Sun et al. (2022) developed a multivariate time-lag sliding window ML framework for forecasting desert locust presence. By integrating NDVI, soil moisture, precipitation, and temperature variables with temporal lags, they demonstrated marked improvements in predictive stability. Gómez et al. (2018) applied machine learning to ESA CCI soil moisture data for identifying desert locust breeding areas, revealing that soil moisture

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

anomalies serve as leading indicators of reproductive site formation.

Lawton et al. (2022) advanced locust ecological modeling by incorporating spatiotemporal hierarchy into predictive frameworks. Their findings showed that accounting for nested spatial structure improves model generalization and reduces prediction bias. This aligns with broader geospatial methodological advances outlined by Goodchild (2009), who formalized spatial data structures and geostatistical modeling within GIS science.

Accuracy assessment remains a critical component of ML–GIS integration. Congalton & Green (2019) established rigorous protocols for evaluating the accuracy of remotely sensed classifications, emphasizing the use of confusion matrices, kappa statistics, and spatial validation procedures. Such evaluation frameworks are particularly important when forecasting outputs are intended to support policy-level phytosanitary surveillance and management decisions. Klein et al. (2021) further highlighted the increasing integration of Earth observation (EO) data into locust monitoring and early-warning systems. Their synthesis demonstrated the value of remotely sensed indicators, including normalized difference vegetation index (NDVI) anomalies, land surface temperature, and soil moisture, for identifying environmental conditions associated with locust population dynamics. Subsequently, Klein et al. (2023) demonstrated the operational integration of EO-derived predictors into geospatial locust management platforms, supporting the feasibility of linking ML-based predictions with spatial decision-support systems.

Collectively, the available literature indicates that contemporary locust forecasting increasingly relies on several complementary methodological components:

1. Nonparametric ensemble algorithms for capturing nonlinear relationships between locust occurrence and environmental predictors (Hastie et al., 2009; de Oliveira Aparecido et al., 2020; Singh et al., 2024);
2. Recurrent neural networks for modeling temporal dependencies and sequential ecological processes (Xiao et al., 2019);
3. Deep-learning architectures for extracting complex patterns from multidimensional environmental and Earth observation data (Reichstein et al., 2019);
4. Lag-integrated ML frameworks for incorporating delayed environmental effects into outbreak prediction (Gómez et al., 2018; Sun et al., 2022);
5. Hierarchical and spatially explicit modeling approaches for accounting for spatial heterogeneity, scale dependence, and geographic structure (Goodchild, 2009; Congalton & Green, 2019; Lawton et al., 2022);
6. Operational ML–GIS integration for translating predictive outputs into spatially explicit early-warning and decision-support systems (Klein et al., 2021, 2023).

Despite these methodological advances, most existing applications have focused on desert locust systems or selected economically important crop pests. Comparable ML–GIS frameworks for forecasting non-gregarious locust populations in continental Eurasian agroecosystems remain comparatively underdeveloped. In particular, there is a need for regionally calibrated approaches capable of integrating multi-source climatic, ecological, and remotely sensed predictors while explicitly accounting for spatial and temporal variability. Addressing this gap provides a strong rationale for developing an ensemble-based phytosanitary forecasting framework integrated within a GIS-based spatial decision-support architecture, with the potential to improve risk mapping, early warning, surveillance prioritization, and targeted locust management at regional scales.

1.4. GIS and Remote Sensing for Spatial Risk Modeling

The transition from descriptive habitat assessment to fully operational forecasting platforms requires integrating machine learning algorithms with spatially explicit GIS infrastructures. Recent empirical studies demonstrate that such integration substantially improves outbreak detection accuracy and spatial prioritization of intervention zones.

Klein et al. (2021) systematically analyzed remote sensing applications in locust ecology and demonstrated that vegetation greening dynamics derived from NDVI time series provide early signals of breeding habitat formation. Their review emphasized the predictive value of MODIS-derived vegetation indices in identifying pre-outbreak ecological conditions. Building on this approach, Klein et al. (2023) implemented a geospatial decision-support framework integrating Earth observation datasets into operational locust management systems. Their study showed that combining satellite-derived vegetation metrics with climatic thresholds allows for spatially prioritized intervention planning, reducing response time during outbreak phases.

Methodologically, Elith & Leathwick (2009) showed that species distribution modeling (SDM) frameworks enable robust estimation of nonlinear environmental response curves. Phillips et al. (2006) formalized the Maximum Entropy (MaxEnt) approach for presence-only ecological data, providing a mathematically consistent solution for habitat suitability modeling under incomplete sampling conditions. Subsequent methodological refinement by Elith et al. (2009) improved interpretability and predictive performance of SDMs through regularization and cross-validation strategies.

In the context of ensemble forecasting, Thuiller et al. (2009) introduced BIOMOD, demonstrating that combining multiple algorithms significantly increases robustness under environmental uncertainty. Araújo & New (2007) further

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

showed that ensemble approaches outperform single-model predictions when projecting species distributions under climate change scenarios. Cross-scale ecological theory supports these modeling strategies. Peters et al. (2009) highlighted that ecosystem processes operate across spatial and temporal scales, necessitating hierarchical modeling structures. Reichstein et al. (2019) demonstrated that deep learning techniques can detect nonlinear patterns in Earth system data beyond the capacity of traditional regression models. Ibrahim et al. (2022) provided an applied example of expert-system integration with machine learning for agricultural pest population forecasting, illustrating operational feasibility. High-resolution climatic datasets, such as WorldClim (Fick & Hijmans, 2017), provide spatially continuous environmental predictor surfaces that are widely used in ecological and species-distribution modeling. Similarly, Shuttle Radar Topography Mission (SRTM)-derived digital elevation models provide detailed information on elevation and terrain variability, thereby improving the topographic characterization and spatial stratification of ecological niches. In parallel, model-interpretability approaches such as SHapley Additive exPlanations (SHAP) enhance the transparency of machine-learning predictions by quantifying the relative contribution of individual predictors and identifying the climatic and environmental variables that most strongly influence predicted outbreak risk.

Within the Central Asian context, Baibussenov et al. (2024) investigated habitat suitability and phytosanitary zoning for harmful non-gregarious locusts. Their findings demonstrated the value of integrating satellite-derived vegetation indicators with agro-climatic information for identifying areas associated with elevated reproductive and phytosanitary risk in northern Kazakhstan. This regional evidence reinforces the importance of spatially explicit modeling approaches that account for local climatic, vegetation, and landscape gradients when assessing non-gregarious locust populations. Despite these advances, the development of integrated ML–GIS approaches for dynamic forecasting of harmful non-gregarious locust populations in Kazakhstan remains limited. Existing regional research has largely emphasized habitat suitability, ecological zoning, and spatial risk identification, whereas comparatively little attention has been given to integrating dynamic population forecasting, multi-source environmental predictors, ensemble machine-learning algorithms, model interpretation, GIS-based risk mapping, and web-enabled decision support within a unified operational framework. This represents an important methodological and applied research gap.

Accordingly, the development of an ensemble-based ML–GIS forecasting framework that integrates multi-source climatic, topographic, vegetation, and other environmental predictors with spatial deployment provides a logical advancement in regional phytosanitary surveillance. Such a framework could move beyond static habitat-suitability assessment toward dynamic and interpretable risk forecasting, enabling the identification of high-risk areas, prioritization of field surveillance, and timelier and spatially targeted management of harmful non-gregarious locust populations in Kazakhstan.

1.5. Research Gap and Objective

Given the pronounced agro-climatic heterogeneity of Kazakhstan and the ecological specificity of harmful non-gregarious locust species, regionally calibrated predictive frameworks are required. Previous GIS-based studies in northern Kazakhstan have demonstrated the feasibility of habitat-suitability mapping; however, important methodological gaps remain. In particular, the integration of nationwide population-density forecasting, temporally independent evaluation of multiple predictive algorithms, agro-climatic risk stratification, and GIS-based translation of model outputs into a unified phytosanitary decision-support framework remains limited. Therefore, this study aimed to develop and validate an integrated Machine Learning–Geographic Information System (ML–GIS) framework for forecasting the population density and phytosanitary risk of harmful non-gregarious locusts across the agro-climatic zones of Kazakhstan. The proposed framework integrates (i) multi-year phytosanitary monitoring data, (ii) climatic and meteorological predictors, (iii) satellite-derived vegetation and soil-moisture indices, and (iv) ensemble and other machine-learning algorithms to generate spatially explicit predictions of locust population density and associated risk. Three hypotheses were tested. H1: Relationships between the absolute population density of harmful non-gregarious locusts and hydrothermal predictors are nonlinear and vary among agro-climatic risk zones. H2: The dominant environmental controls on locust population density differ among risk zones: high-risk zones are primarily influenced by thermal accumulation and surface-energy conditions, moderate-risk zones by interactions between temperature and moisture, and low-risk zones by limiting thermal or moisture conditions. Accordingly, population density is expected to increase under warm and moderately moist conditions but decline under extreme heat, drought, excessive soil moisture, or insufficient thermal resources. H3: Nonlinear ensemble and kernel-based machine-learning models achieve lower out-of-sample prediction errors than linear regression and conventional time-series approaches, with the best-performing algorithm varying among agro-climatic risk zones. By integrating ecological predictors, population monitoring, predictive modeling, and spatial risk mapping within a unified framework, this study seeks to support a transition from predominantly reactive locust control toward proactive, risk-based phytosanitary surveillance and population management in Kazakhstan.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

2. MATERIALS AND METHODS

2.1. Study Area

The study covered the territory of the Republic of Kazakhstan, extending approximately from 40° to 56° N and from 46° to 88° E. Kazakhstan encompasses a pronounced continental environmental gradient from forest-steppe and steppe landscapes in the north to dry steppe, semi-desert, and desert landscapes in the central, western and southern parts of the country, as well as mountain forest, meadow, and alpine belts in the east and southeast. Annual temperature amplitudes commonly exceed 35–40°C, while long-term annual precipitation generally ranges from approximately 150 mm in the desert regions to more than 400 mm in the northern and mountainous regions (Mingalev, 2021).

The study area was divided into 25 agroclimatic zones according to two principal characteristics: moisture availability and thermal regime. The zones were consolidated into eight major moisture groups: extremely arid, highly arid, moderately arid, slightly humid, sufficiently humid, persistently humid, optimally humid, and extremely humid. Each moisture group included one or more thermal subclasses ranging from slightly warm to hot. The numerical codes, complete names, monitoring periods, numbers of observations, and long-term population statistics of all agroclimatic zones are presented in Table 1. The monitoring territory included the West Kazakhstan, Atyrau, Mangystau, Aktobe, Kostanay, North Kazakhstan, Akmola, Pavlodar, Karaganda, Ulytau, Kyzylorda, Turkistan, Zhambyl, Almaty, Zhetysu, Abai, and East Kazakhstan regions. Representative regional reference points shown in Fig. 1 include western Kazakhstan near Atyrau (47.10° N, 51.92° E), northern Kazakhstan near Kostanay (53.21° N, 63.62° E), central Kazakhstan near Karaganda (49.80° N, 73.10° E), southern Kazakhstan near Kyzylorda (44.85° N, 65.50° E), southeastern Kazakhstan near Almaty (43.24° N, 76.95° E), and eastern Kazakhstan near Oskemen (49.95° N, 82.63° E). These coordinates are regional reference points and should not be interpreted as the exact coordinates of every historical field survey because the long-term monitoring data were predominantly aggregated by administrative unit and agroclimatic zone. All spatial data were harmonized in the World Geodetic System 1984 coordinate reference system (WGS 84; EPSG: 4326). The revised location map presents the position of Kazakhstan within Central Asia, national and regional boundaries, the 25 agroclimatic zones, representative monitoring regions, latitude and longitude, a north arrow, scale bar and an expanded legend.

Table 1: Statistics of the number of observations and the mean absolute infestation index (P_{abs}, pcs./m²) within the agroclimatic zones of Kazakhstan (2003–2024)

#	Agroclimatic Zone	Number of Observations	Mean Pabs (pcs./m²)	Median	Min	Max	Std. Dev.
I. Extremely arid zones							
1	(15) Extremely arid / Hot	87	2.91	0.429	0.00	155.5	16.70
II. Highly arid zones							
2	(22) Highly arid / Moderately warm	65	3.359	1.100	0.000	52.776	8.907
3	(23) Highly arid / Warm	455	23.1	0.540	0.00	9955.0	466.6
4	(24) Highly arid / Moderately hot	939	1.507	0.438	0.000	155.55	7.864
5	(25) Highly arid / Hot	224	2.72	1.500	0.00	68.78	6.326
III. Moderately arid zones							
6	(32) Moderately arid / Moderately warm	512	1.84	0.81	0.0	38.45	3.21
7	(33) Moderately arid / Warm	723	1.75	0.735	0.00	43.69	3.093
8	(34) Moderately arid / Moderately hot	76	1.96	0.922	0.00	15.32	2.636
9	(35) Moderately arid / Hot	687	2.39	0.516	0.00	474.7	19.14
IV. Slightly humid zones							
10	(41) Slightly humid / Slightly warm	356	2.98	1.02	0.0	112.4	9.87
11	(42) Slightly humid / Moderately warm	923	12.455	1.352	0.000	9436.275	310.687
12	(44) Slightly humid / Moderately hot	478	3.65	1.24	0.0	186.75	12.54
13	(45) Slightly humid / Hot	401	1.69	0.829	0.00	86.78	4.815
V. Sufficiently humid zones							
14	(51) Sufficiently humid / Slightly warm	66	1.951	0.994	0.000	11.667	2.326
15	(52) Sufficiently humid / Moderately warm	75	1.997	1.035	0.000	13.411	2.544
16	(54) Sufficiently humid / Moderately hot	58	1.73	0.88	0.0	9.85	2.11
17	(55) Sufficiently humid / Hot	32	1.092	0.503	0.000	6.645	1.643
VI. Persistently humid zones							
18	(61) Persistently humid / Slightly warm	142	1.62	0.72	0.0	21.4	3.34
19	(65) Persistently humid / Hot	169	1.99	0.900	0.00	36.99	4.051
VII. Optimally humid zones							
20	(71) Optimally humid / Slightly warm	240	1.275	0.550	0.000	11.211	1.699
21	(75) Optimally humid / Hot	65	8.057	1.871	0.000	343.40	42.38
VIII. Extremely humid zones							
22	(81) Extremely humid / Slightly warm	74	1.41	0.63	0.0	12.75	2.26
23	(82) Extremely humid / Moderately warm	88	1.513	0.558	0.000	11.530	2.124
24	(83) Extremely humid / Warm	88	1.99	1.010	0.00	15.32	2.548
25	(85) Extremely humid / Hot	169	3.30	1.500	0.00	183.9	14.33

Note: P_{abs} is expressed as individuals m^{-2}; n, number of monitoring records; SD, standard deviation. The median and full range are presented because several zone-specific distributions were strongly right-skewed.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

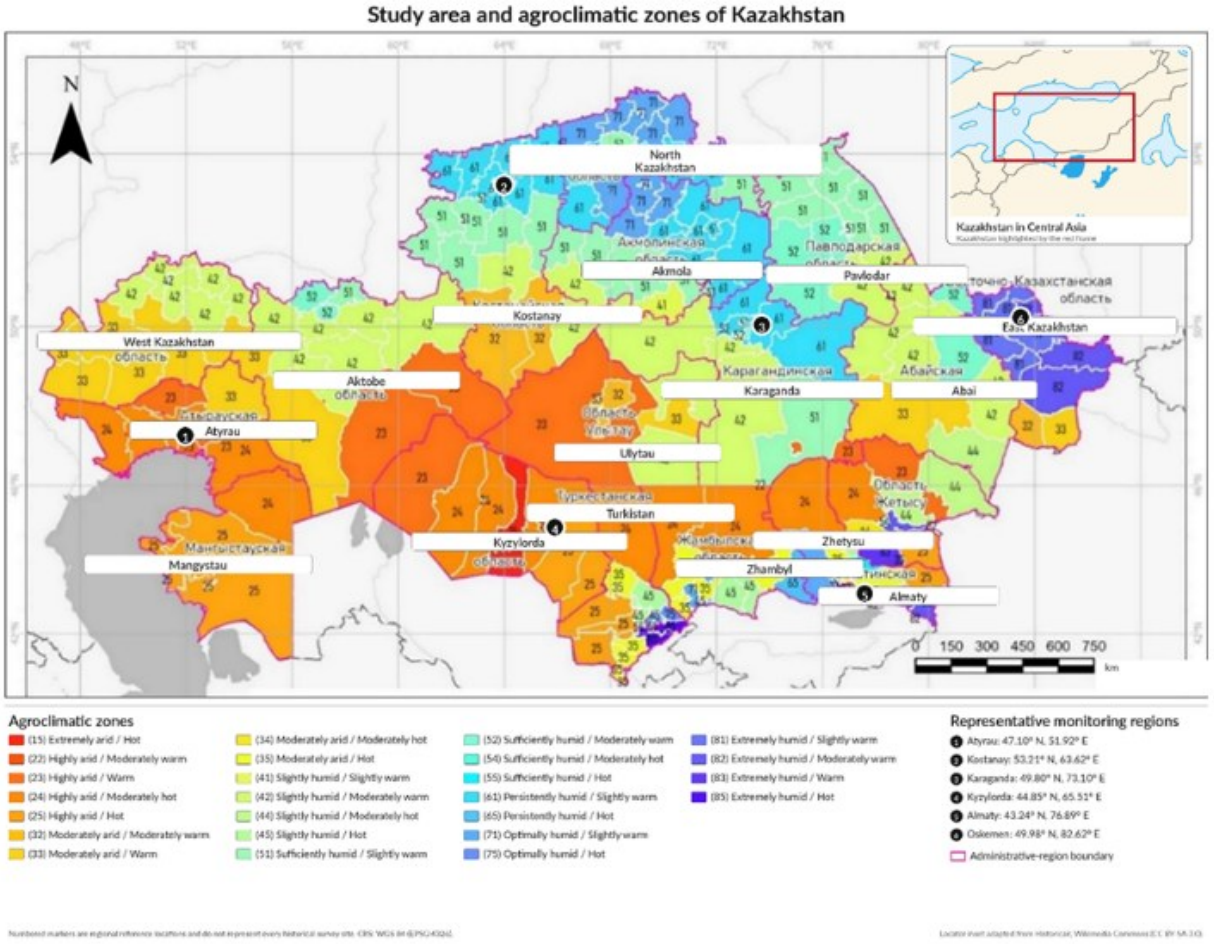

Fig. 1: Agroclimatic zones of Kazakhstan.

2.2. Climatic Conditions During the Study Period

The study period (2003–2024) coincided with a pronounced warming trend in Kazakhstan. According to national meteorological observations, mean annual air temperature increased by approximately 0.36°C per decade during 1976–2024**.** Several exceptionally warm years occurred within the study period. In 2023, the nationally averaged annual air-temperature anomaly reached +2.58°C relative to the 1961–1990 reference period**,** making it the warmest year recorded in Kazakhstan since national observations began in 1941. In 2024, the corresponding anomaly was +1.72°C, ranking the year sixth among the warmest on record. Precipitation exhibited substantially greater spatial and interannual variability than temperature. In 2024, nationally averaged annual precipitation reached 391.8 mm, equivalent to 123.3% of the climatic norm**,** although precipitation deficits occurred simultaneously in parts of northwestern, central, and southern Kazakhstan. Over 1976–2024, the national trend in annual precipitation was weak and statistically non-significant, whereas air temperature showed a persistent increase across the major regions of the country. Consequently, the 2003–2024 study period encompassed contrasting warm, cool, dry and wet conditions rather than a uniform climatic regime**.** Such hydroclimatic variability is particularly relevant to locust population dynamics because temperature and moisture conditions can influence development, reproduction, vegetation availability, and habitat suitability. Therefore, temporally resolved predictors—including temperature, precipitation, soil moisture, evaporation, and vegetation indices—were incorporated into the modeling framework rather than relying exclusively on long-term climatic averages.

2.3. Data Sources

2.2.1. Phytosanitary Monitoring Data: Long-term historical data on the distribution and abundance of harmful non-gregarious locusts were obtained from the state phytosanitary monitoring services of Kazakhstan for 2003–2024. The compiled database contained 7192 monitoring records distributed among the 25 agroclimatic zones. The available fields included the year and survey period, administrative region, agroclimatic-zone code, surveyed area, populated area, population-density class, mean density within the populated area, and, where available, the

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

developmental stage of the insects. The historical observations represented the economically important complex of harmful non-gregarious acridid species rather than a single species. State monitoring was conducted according to the nationally adopted phytosanitary survey and pest-accounting procedures, including route observations, visual density assessment, and sweep-net sampling in representative biotopes. Because operational phytosanitary thresholds and protective decisions are applied to the combined harmful species complex, the population response used in this study was calculated for the entire complex. Before analysis, every record was assigned to one of the 25 agroclimatic zones. Data were subsequently analysed both by individual agroclimatic zone and by the three reproduction and harmfulness risk categories.

2.2.2. Calculation of Population Indices: Three population indicators were calculated: relative population (P_{rel}), average population density within the populated area (P_{av}) and absolute population density (P_{abs}).

Relative population was calculated as:

$$P_{rel} = \frac{A_p}{A_s} \times 100$$

Where P_{rel} is relative population, %;
A_p is the populated area, thousand ha;
And A_s is the surveyed area, thousand ha.

Average population density within the populated area was calculated as an area-weighted mean across the density classes:

$$P_{\omega} = \frac{\sum_{k=1}^{k} A_{P,k} N_{av,k}}{\sum_{k=1}^{K} A_{P,k}}$$

Where P_a is average population density, individuals m^{-2};
$A_{p,k}$ is the populated area assigned to density class k, thousand ha;
$N_{av,k}$ is the mean population density within class k, individuals m^{-2};
And K is the number of population-density classes.

Absolute population density was calculated as:

$$P_{abs} = \frac{P_{rel} \times P_{av}}{100}$$

Where P_{abs} is absolute population density, individuals m^{-2}.

The P_{abs} index adjusts the density recorded in occupied sites for the proportion of the total surveyed territory that was populated. It was used as the response variable in all statistical and machine-learning models. Initial calculation and verification of P_{rel}, P_{av}, and P_{abs} were performed in Microsoft Excel. No neural-network model was applied in the present study

2.2.3. Meteorological and Climatic Data: Meteorological predictors included mean, minimum, and maximum air temperature; land-surface temperature; Growing Degree Days (GDD); monthly precipitation; relative humidity; evaporation; solar radiation; and soil moisture. The climatic variables were obtained from the national meteorological monitoring system and global climatic databases. WorldClim version 2 was used to characterize the long-term spatial climatic background, whereas year-specific meteorological observations were used to represent interannual variability during 2003–2024 (Fick & Hijmans, 2017).

Predictors were compiled primarily for April–October, covering the biologically active period that includes egg hatching, larval development, adult activity, reproduction, and oviposition. Monthly values were matched to the corresponding year, monitoring period, administrative unit, and agroclimatic zone. Depending on the predictor, monthly means, monthly totals, lagged values, and moving averages were calculated.

Growing Degree Days were calculated as the cumulative positive difference between daily mean temperature and a base temperature of 10°C:

$$GDD = \sum_{d=1}^{D} max\left(0, T_{mean,d} - 10\right)$$

Where $T_{mean,d}$ is the mean air temperature on day d, °C, and D is the number of days in the corresponding accumulation period.

The Standardized Precipitation Index (SPI) was used as a dimensionless indicator of precipitation conditions relative to the long-term distribution for the corresponding period. Values close to zero represented near-normal precipitation, negative values represented drier-than-normal conditions, and positive values represented wetter-than-normal conditions.

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

2.2.4. Remote-Sensing and Topographic Predictors: The remote-sensing predictors included the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), land-surface temperature (LST), Leaf Area Index (LAI), and satellite-derived soil-moisture indicators. MODIS data were used as the principal temporally consistent source of vegetation and land-surface variables over the full study period. Sentinel-2 multispectral imagery, Landsat 8/9 thermal data, and SMAP/SMOS soil-moisture products were used as supplementary sources for the years in which they were available and for spatial verification of recent conditions (Didan et al., 2019). They were not treated as a single continuous sensor record covering all years since 2003.
Vegetation indices were calculated using the standard equations:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

$$EVI = 2.5 \times \frac{NIR - RED}{NIR + 6RED - 7.5BLUE + 1}$$

Where *NIR*, *Red*, and *Blue* represent atmospherically corrected surface reflectance in the near-infrared, red, and blue spectral bands, respectively (Didan et al., 2019). Satellite observations were summarized by monitoring period and agroclimatic zone. Topographic variables, including elevation and slope, were derived from Shuttle Radar Topography Mission digital-elevation data.

2.2.5. Predictor Definitions and Feature Engineering: The response variable was absolute population density (Pabs, individuals m^{-2}). In the analytical dataset and graphical outputs, the internal variable name quantity denotes Pabs. The basic environmental predictors included mean air temperature (AirTemp, °C), land-surface temperature (LST, °C), precipitation (mm), relative humidity (%), soil moisture, evaporation, solar radiation (SolarRad), Growing Degree Days (GDD, °C·day), the Standardized Precipitation Index (SPI, dimensionless), the Normalized Difference Vegetation Index (NDVI, dimensionless), the Enhanced Vegetation Index (EVI, dimensionless) and the Leaf Area Index (LAI, $m^2\ m^{-2}$). Soil-moisture values used in the models were transformed to a normalized 0–1 scale using Min–Max normalization.

Derived predictors were calculated to represent nonlinear and combined environmental effects. TempMoisture_interaction was the product of air temperature and soil moisture; PrecipSPI_interaction was the product of precipitation and SPI; and LSTEvap_interaction was the interaction between land-surface temperature and evaporation. AirTemp_squared, Precipitation_squared, and SoilMoisture_squared were quadratic terms used to represent possible nonlinear responses. Aridity_Index represented the ratio of evaporation to precipitation, whereas Moisture_Index represented the inverse moisture–aridity relationship. Heat_Stress_Index was calculated as the mean of LST and air temperature and Temp_Amplitude as the difference between LST and air temperature. Energy_Balance was an empirical derived index representing the difference between the solar-radiation and evaporation variables used in the model.

The variable Season_numeric encoded spring as 1, spring–summer as 1.5, and summer as 2. Prev_quantity, Prev_precipitation, and Prev_temperature represented the corresponding variables from the preceding monitoring period. Quantity_ma3 and Temp_ma3 represented trailing three-period moving averages of Pabs and air temperature, respectively. Year_trend was a sequential temporal variable used to represent long-term change from the beginning of the study period. All lagged and moving-average predictors were calculated within the corresponding agroclimatic zone using only preceding observations.

2.4. Data Cleaning, Missing-Value Handling and Feature Engineering

Before modelling, the combined database was checked for duplicate records, inconsistent agroclimatic-zone codes, impossible dates, negative values of surveyed or populated area, populated area exceeding surveyed area, and values outside the physically or biologically plausible ranges of the corresponding variables. Exact duplicate records were removed. Suspected data-entry or coding errors were verified against the original phytosanitary records whenever these were available.

Potential outliers were identified separately within the risk categories using the 1.5 × interquartile-range criterion. This procedure was used to flag observations for verification rather than to remove them automatically. High P_{abs} values were retained when they represented documented population outbreaks because such extreme observations constituted an essential part of the ecological process being modelled. Only records judged to represent data-entry, coding, or measurement errors were excluded. Missing values were quantified separately for every predictor. Missing values of the response variable P_{abs} were not imputed; records lacking a valid response value were excluded from model fitting. Missing predictor values were imputed using the k-nearest-neighbour procedure. The imputation model was fitted to the model-development dataset and subsequently applied to the independent test dataset so that information from 2019–2024 did not influence the preprocessing of earlier observations.

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

Multicollinearity among continuous predictors was assessed using the Variance Inflation Factor (VIF). A VIF threshold of 5 was applied. When two or more variables contained strongly redundant information, the variable with clearer ecological interpretation or more complete temporal coverage was retained. VIF-based predictor reduction was applied primarily to Multiple Linear Regression, whereas tree-based models were allowed to evaluate correlated predictors through their internal split-selection procedures.

Continuous predictors used by scale-sensitive algorithms were transformed to the interval 0–1 using Min–Max normalization. The minimum and maximum values used for normalization were estimated from the 2003–2018 model-development dataset and then applied unchanged to the 2019–2024 test dataset. P_{abs} was evaluated in its original unit of individuals m^{-2} so that MAE and RMSE remained biologically interpretable. Temporal predictors were constructed using one- to three-period lags. Lagged variables were calculated within the same agroclimatic zone, and only observations preceding the predicted period were used. Three-period moving averages were calculated as trailing averages and did not include future values. Polynomial variables and interaction terms were created to represent nonlinear and combined hydrothermal effects. Exploratory data analysis was then used to identify response patterns and environmental ranges associated with increased or reduced population density. The resulting hydrothermal ranges were interpreted as descriptive ecological indicators rather than as independently validated universal physiological thresholds.

2.5. Classification of Reproduction and Harmfulness Risk

The 25 agroclimatic zones were stratified into three reproduction and harmfulness risk categories according to their mean P_{abs} for 2003–2024. The classification was used as an ecological stratification of the monitoring territory rather than as a predicted response. High risk was defined as mean P_{abs}>8.0 individuals m^{-2}, moderate risk as mean P_{abs}=2.0–8.0 individuals m^{-2}, and low risk as mean P_{abs}<2.0 individuals m^{-2}. The high-risk category included zones (23) Highly arid / Warm, (42) Slightly humid / Moderately warm, and (75) Optimally humid / Hot. The moderate-risk category included zones (15) Extremely arid / Hot, (22) Highly arid / Moderately warm, (25) Highly arid / Hot, (35) Moderately arid / Hot, (41) Slightly humid / Slightly warm, (44) Slightly humid / Moderately hot, and (85) Extremely humid / Hot.

The low-risk category included zones (24) Highly arid / Moderately hot, (32) Moderately arid / Moderately warm, (33) Moderately arid / Warm, (34) Moderately arid / Moderately hot, (45) Slightly humid / Hot, (51) Sufficiently humid / Slightly warm, (52) Sufficiently humid / Moderately warm, (54) Sufficiently humid / Moderately hot, (55) Sufficiently humid / Hot, (61) Persistently humid / Slightly warm, (65) Persistently humid / Hot, (71) Optimally humid / Slightly warm, (81) Extremely humid / Slightly warm, (82) Extremely humid / Moderately warm, and (83) Extremely humid / Warm. Models were fitted and evaluated separately for the three risk categories because the hypotheses predicted that the relative effects of temperature, moisture, and their interactions would differ along the national hydrothermal gradient.

2.6. Machine-Learning and Time-Series Modelling

Five forecasting approaches were compared: Multiple Linear Regression (MLR), Support Vector Regression (SVR), Seasonal Autoregressive Integrated Moving Average (SARIMA), Random Forest Regression (RF), and Gradient Boosting Regression (GBR). The response variable was Pabs, expressed as individuals m^{-2}, and the predictor matrix contained the climatic, remote-sensing, temporal, polynomial, and interaction variables defined in Table 3.

Multiple Linear Regression was included as an interpretable parametric baseline. To reduce overfitting and multicollinearity, the linear model was restricted to three or four principal predictors within each risk category after correlation and VIF screening. Support Vector Regression was used to model smooth nonlinear relationships between Pabs and the environmental predictors. A radial-basis-function kernel was applied. Predictor normalization was performed before fitting SVR, and all model fitting was restricted to the model-development period. SARIMA was included as a time-series benchmark to represent autoregressive and moving-average dependence. The non-seasonal parameters p, d, and q, and, where supported by the temporal structure, the seasonal parameters P, D, Q, and s, were evaluated using the training series. The final specification was selected according to the lowest Akaike Information Criterion and forecast stability within the model-development period. Random Forest Regression was implemented using 100 regression trees and a maximum tree depth of 3. Gradient Boosting Regression was implemented using 100 boosting stages, a maximum tree depth of 3, and a learning rate of 0.1. These deliberately restricted tree depths were used to reduce overfitting in the temporally structured dataset and were consistent with the authors' preceding regional modelling framework (Baibussenov et al., 2025). Model tuning was conducted using only the 2003–2018 model-development data. The independent 2019–2024 observations were not used to select predictors, define normalization parameters, fit the missing-value imputer, select SARIMA orders, or choose the final model. An exhaustive random hyperparameter search was not applied; instead, a restricted parameterization was used to ensure comparability among risk categories and to limit overfitting.

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

2.6.1. Training, Testing, and Validation Strategy: The data were divided chronologically rather than randomly. Observations from 2003 through 2018 formed the model-development dataset, while observations from 2019 through 2024 formed an independent out-of-time test dataset. The former statement that this represented an 80/20 split was removed because, when expressed by years, the division comprised 16 development years and six test years. All preprocessing steps, including missing-value imputation, predictor selection, VIF assessment, and Min–Max normalization, were fitted using the 2003–2018 data and then applied to the 2019–2024 data. The test dataset was evaluated only after the modelling procedure had been finalized. Conventional random k-fold cross-validation was not used because random redistribution of observations would mix earlier and later years and could produce temporal information leakage. Instead, model validation was based on the fixed chronological holdout period of 2019–2024. This procedure tested the ability of each algorithm to predict years that occurred after the complete model-development period. Models were fitted separately for the high-, moderate-, and low-risk categories.

2.6.2. Forecasting for 2025–2030: After test-period evaluation, the model selected for each risk category was refitted using the available historical series and used to generate forecasts for 2025–2030. Future estimates were produced under a continuation scenario in which the climatic predictors and their lagged or moving-average values remained within the range observed during the recent monitoring period, without imposing an external extreme-climate scenario. Consequently, the 2025–2030 estimates represent conditional projections based on the continuation of recently observed environmental and population patterns rather than deterministic predictions under a specified CMIP climate-change scenario.

2.6.3. Model Evaluation and Predictor Analysis: Predictive performance in the independent 2019–2024 test period was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where yi is the observed P_{abs}, $\hat{y}i$ is the predicted P_{abs}, $\bar{y}$ is the mean observed value, and n is the number of observations in the test dataset. MAE and RMSE were expressed as individuals m^{-2}.

RMSE was used as the primary model-selection criterion because large forecast errors during population increases are operationally more consequential than small deviations during background-density periods. MAE was used as a complementary measure of the typical absolute prediction error, and R^2 indicated performance relative to prediction based on the mean observed test value. Negative R^2 values were interpreted as performance below the mean-only test-set benchmark and not as evidence of strong explanatory capacity. Pearson correlation coefficients were calculated to assess the direction and strength of pairwise relationships between each predictor and P_{abs}. The 15 predictors with the highest absolute correlation coefficients were presented separately for the high-, moderate-, and low-risk categories. Therefore, Fig. 2–4 represent predictor–target correlation structures and not Random Forest feature-importance values. Correlation results were treated as exploratory associations and were not interpreted as proof of causality.

Potential interannual outbreak intervals were assessed descriptively from the annual observed population series by comparing the timing of the principal population peaks. Because the 22-year observation period contains only a limited number of major peaks, the resulting intervals were interpreted as apparent or quasi-cyclic patterns rather than as formally established periodicities based on spectral or wavelet analysis.

2.6.4. GIS Integration: GIS was used to assign historical records to administrative regions and agroclimatic zones, visualize the national hydrothermal gradient, and link the model outputs to the three phytosanitary risk categories. Observed and predicted P_{abs} values were joined to the corresponding agroclimatic-zone codes. The GIS component provided the spatial framework for risk stratification and decision support. Fig. 1 presents the geographic distribution of the agroclimatic zones and representative monitoring regions, whereas Fig. 5–7 present the temporal trajectories of observed and predicted population density for the three risk categories. The latter figures should therefore be interpreted as temporally stratified forecasting outputs linked to GIS-defined zones rather than as raster risk maps.

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

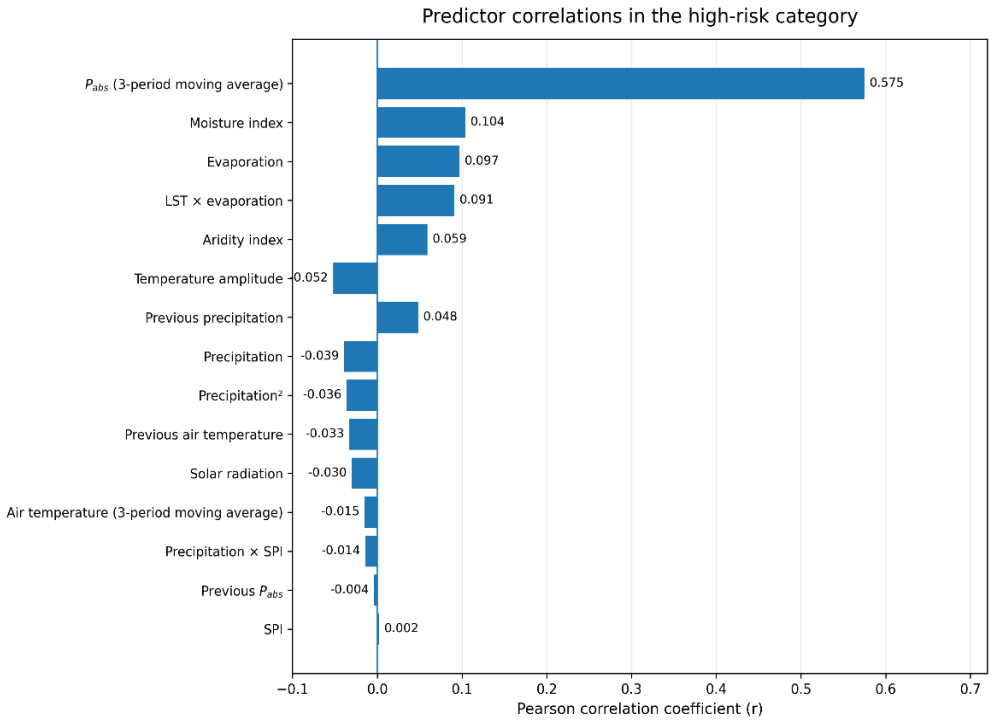

Fig. 2: Pearson correlations between absolute population density and the 15 highest-ranked predictors in the high-risk category. Bars indicate Pearson correlation coefficients (r).

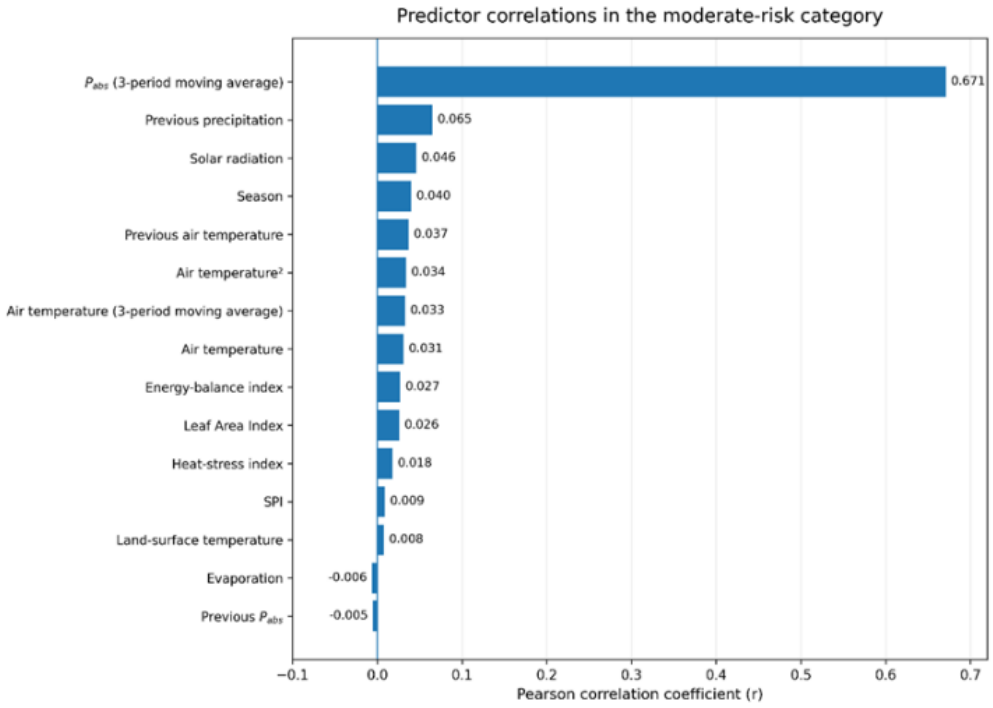

Fig. 3: Pearson correlations between absolute population density and the 15 highest-ranked predictors in the moderate-risk category. Bars indicate Pearson correlation coefficients (r).

2.7. Software

Calculation and initial verification of the population indices were performed in Microsoft Excel. Data cleaning, feature engineering, statistical analysis, and modelling were conducted in the Google Colab cloud environment using Python-based tools. The principal libraries included pandas and NumPy for data management, scikit-learn for Multiple Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, scaling, imputation, and regression metrics, statsmodels for SARIMA modelling, and Matplotlib for graphical presentation. GIS processing and preparation of the agroclimatic-zone map were performed using geospatial software in the WGS 84 coordinate reference system.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

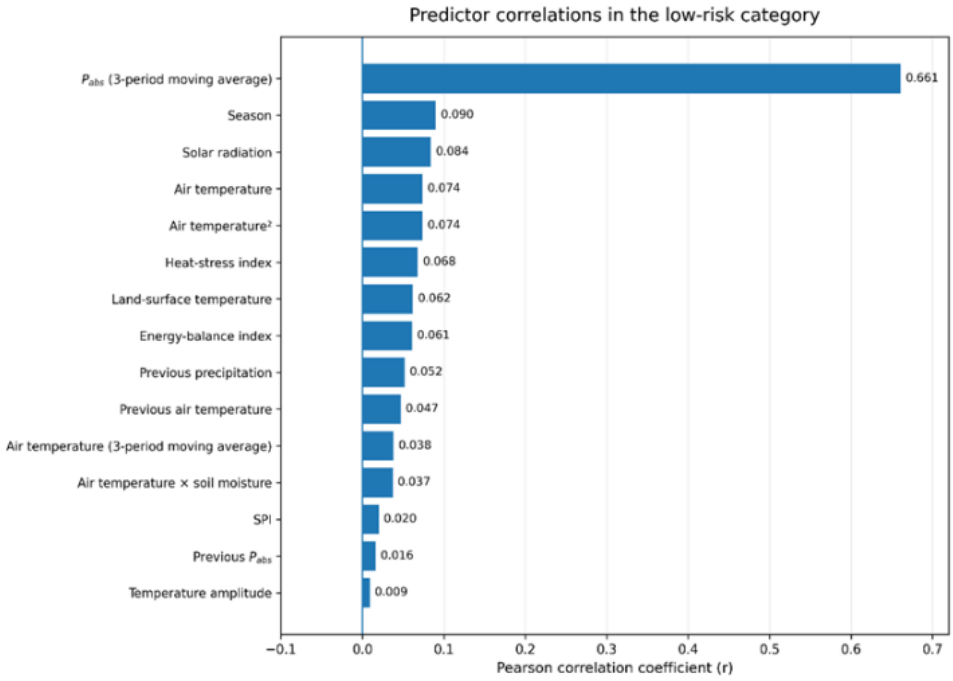

Fig. 4: Pearson correlations between absolute population density and the 15 highest-ranked predictors in the low-risk category. Bars indicate Pearson correlation coefficients (r).

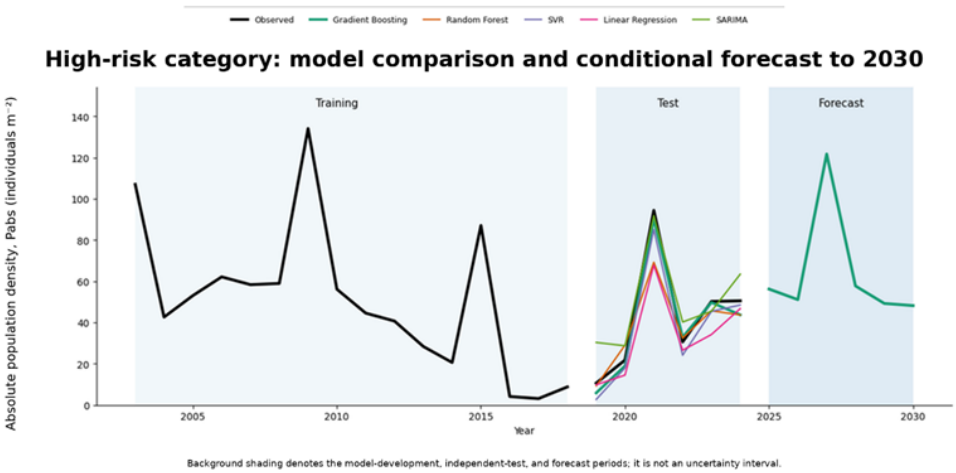

Fig. 5: Observed and model-predicted absolute population density in the high-risk category during the model-development period (2003–2018), independent test period (2019–2024), and conditional forecast period (2025–2030). Background shading denotes the temporal partition and does not represent a confidence or prediction interval.

3. RESULTS

3.1. Spatial Variation in Population Density and Risk Classification

The compiled dataset contained 7,192 phytosanitary monitoring records distributed among 25 agroclimatic zones. Mean absolute population density (P_{abs}) varied from 1.092 individuals m^{-2} in the Sufficiently humid / Hot zone (55) to 23.100 individuals m^{-2} in the Highly arid / Warm zone (23) (Table 1). The three zones classified as high risk had the largest long-term mean values but also the greatest within-zone variability. In zone (23), mean P_{abs} was 23.100 individuals m^{-2}, whereas the median was only 0.540 individuals m^{-2}, the standard deviation was 466.600 individuals m^{-2}, and the maximum reached 9,955.000 individuals m^{-2}. Corresponding values were 12.455, 1.352, 310.687, and 9,436.275 individuals m^{-2} in zone (42), and 8.057, 1.871, 42.380, and 343.400 individuals m^{-2} in zone (75), respectively. The large differences between the means and medians indicate strongly right-skewed distributions in which long-term means were influenced by infrequent extreme-density observations rather than by consistently high population levels. Descriptive uncertainty and within-zone variability are therefore presented in Table 1 using the standard deviation, median, and full observed range. Based on the predefined long-term mean P_{abs} thresholds, three agroclimatic zones were assigned to the high-risk category, seven to the moderate-risk category, and 15 to the low-risk category (Table 2). The moderate-risk group included zones (15), (22), (25), (35), (41), (44), and (85), with mean P_{abs} values ranging from 2.390 to 3.650 individuals m^{-2}. Mean values in the low-risk group ranged from 1.092 to 1.997 individuals m^{-2}. This classification was used as the spatial stratification framework for the subsequent predictor and forecasting analyses.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

Table 2: Agroclimatic risk classification based on mean absolute population density (P_{abs}), 2003–2024

Risk Level	Mean Pabs (pcs./m²)	Included Agroclimatic Zones
High risk of reproduction and harmfulness	> 8.0	(23); (42); (75)
Moderate risk of reproduction and harmfulness	2.0 – 8.0	(15); (22); (25); (35); (41); (44); (85)
Low risk of reproduction and harmfulness	< 2.0	(24); (32); (33); (34); (45); (51); (52); (54); (55); (61); (65); (71); (81); (82); (83)

Table 3: Out-of-time predictive performance of the tested models in the 2019–2024 test period

Risk category	Model	MAE	RMSE	R^2
High	Gradient Boosting	1.661	2.636	0.005
	SVR	1.307	2.644	-0.002
	SARIMA	143.748	153.630	-3380.069
	Linear Regression	155.025	184.883	-4895.603
	Random Forest	64.075	199.980	-5727.930
Moderate	SVR	1.135	3.039	-0.023
	SARIMA	1.893	3.324	-0.223
	Linear Regression	2.836	3.833	-0.627
	Gradient Boosting	2.151	7.444	-5.137
	Random Forest	2.308	8.661	-7.306
Low	SVR	1.212	1.940	-0.010
	SARIMA	1.762	2.096	-0.178
	Random Forest	1.670	2.147	-0.236
	Gradient Boosting	1.607	2.199	-0.297
	Linear Regression	1.868	2.214	-0.315

Note: MAE and RMSE are expressed as individuals m^{-2}. Bold values indicate the best result within each risk category for the corresponding metric. A negative R^2 indicates performance below the mean-only benchmark for the independent test dataset.

3.2. Predictor–Target Relationships

Fig. 2–4 present Pearson correlations between P_{abs} and the 15 predictors with the largest absolute correlation coefficients in each risk category. These figures represent predictor–target correlation structures and not Random Forest feature-importance rankings.

The trailing three-period moving average of population density (quantity_ma3) was the strongest correlate of current P_{abs} in all three risk categories, with r=0.575 in the high-risk group, r=0.671 in the moderate-risk group, and r=0.661 in the low-risk group (Fig. 2–4). This consistent result indicates that recent population history was more strongly associated with current population density than any individual environmental variable. In the high-risk group, the strongest environmental correlations were observed for Moisture_Index (r=0.104), evaporation (r=0.097), and LSTEvap_interaction (r=0.091), whereas Temp_Amplitude showed a weak negative association (r=−0.052) (Fig. 2). In the moderate-risk group, the largest environmental coefficients were recorded for previous-period precipitation (r=0.065), solar radiation (r=0.046), and season (r=0.040) (Fig. 3). In the low-risk group, season (r=0.090), solar radiation (r=0.084), air temperature (r=0.074), and squared air temperature (r=0.074) had the largest environmental correlations (Fig. 4).

Apart from quantity_ma3, all pairwise coefficients were weak, with absolute values not exceeding 0.104. Thus, the correlation analysis did not identify a single dominant climatic predictor. Rather, it suggests that climatic effects may operate through nonlinear responses, interactions, and temporal lags that cannot be represented fully by individual bivariate correlations. Exploratory response screening indicated that comparatively elevated P_{abs} values were associated with land-surface temperatures of 30–35°C, precipitation totals of 20–50 mm, normalized soil-moisture values of 0.60–0.75, and SPI values of 0–0.5. These values should be interpreted as descriptive hydrothermal ranges identified within the present dataset rather than as independently validated physiological thresholds or universal breakpoint values.

3.3. Model Performance and Conditional Forecasts

Fig. 5–7 compare the observed population trajectories with the outputs of the five tested models during the model-development period of 2003–2018 and the independent test period of 2019–2024. They also show the conditional 2025–2030 projections generated by the selected model for each risk category. Model-performance metrics are summarized in Table 3.

3.3.1. High-Risk Category: The high-risk category exhibited the greatest amplitude of interannual population variation (Fig. 5). Pronounced population maxima occurred approximately in 2003, 2009, 2015, and 2021. Their spacing suggests an apparent inter-peak interval of approximately six years; however, this pattern should be interpreted descriptively because no spectral or wavelet periodicity test was performed.

Gradient Boosting produced the lowest test-period RMSE (2.636 individuals m^{-2}) and the highest R^2 value (0.005), whereas SVR produced the lowest MAE (1.307 individuals m^{-2}) (Table 3). The difference in RMSE between Gradient Boosting and SVR was only 0.008 individuals m^{-2}, indicating that their error-based performance was nearly equivalent. Moreover, the near-zero R^2 values show that neither model explained a substantial proportion of out-of-time population variance. Gradient Boosting was therefore selected on the basis of its marginally lower RMSE rather than on evidence of high predictive accuracy. The conditional Gradient Boosting projection shows a pronounced population increase in 2027 followed by a return toward lower values through 2030 (Fig. 5). The timing of this projected increase is consistent with the apparent six-year spacing of the preceding population maxima, but it should not be interpreted as a deterministic outbreak prediction.

3.3.2. Moderate-Risk Category: The moderate-risk trajectory showed lower overall amplitude than the high-risk trajectory. The principal maxima occurred approximately in 2003, 2011, and 2019, corresponding to an apparent

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

inter-peak spacing of about eight years (Fig. 6). Population density subsequently declined during most of the 2019–2024 test period. SVR produced the lowest MAE (1.135 individuals m^{-2}) and RMSE (3.039 individuals m^{-2}) among the tested algorithms (Table 3). Its R^2 value was nevertheless negative (−0.023-0.023−0.023), indicating that the test-period predictions did not improve on a mean-only benchmark in terms of explained variance. SVR was selected because it minimized the absolute and squared prediction errors relative to the other tested algorithms, not because it demonstrated strong explanatory performance. Contrary to a stable-density interpretation, the SVR projection in Fig. 6 shows a marked increase in 2027, followed by a decline and partial recovery toward 2030. The projected 2027 maximum is consistent with the apparent approximately eight-year spacing of the earlier maxima, although the negative test-period R^2 indicates substantial uncertainty in the magnitude of this projection.

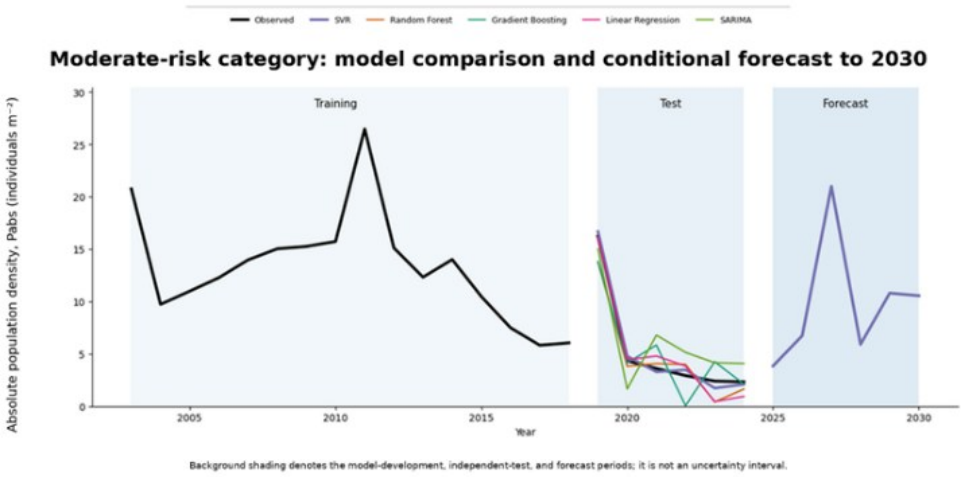

Fig. 6: Observed and model-predicted absolute population density in the moderate-risk category during the model-development period (2003–2018), independent test period (2019–2024), and conditional forecast period (2025–2030). Background shading denotes the temporal partition and does not represent a confidence or prediction interval.

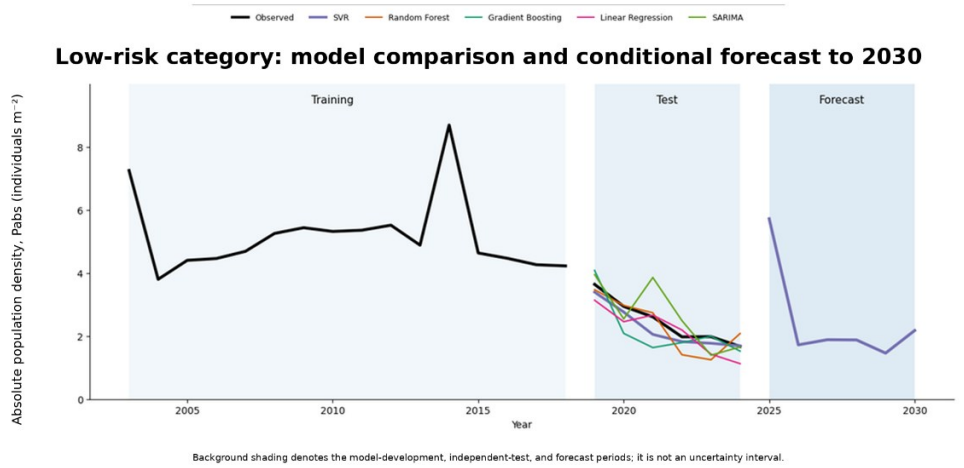

Fig. 7: Observed and model-predicted absolute population density in the low-risk category during the model-development period (2003–2018), independent test period (2019–2024), and conditional forecast period (2025–2030). Background shading denotes the temporal partition and does not represent a confidence or prediction interval.

3.3.3. Low-Risk Category: The low-risk category had the lowest population amplitude. The main observed maxima occurred approximately in 2003 and 2014, suggesting an apparent interval of about 11 years (Fig. 7). Population density declined during most of the 2019–2024 test period and remained below the earlier training-period maximum. SVR had the lowest MAE (1.212 individuals m^{-2}), the lowest RMSE (1.940 individuals m^{-2}), and the highest R^2 value among the tested models (Table 3). However, R^2 remained slightly negative (−0.010-0.010−0.010), showing that the model provided limited out-of-time variance explanation despite having the smallest prediction errors within this category. The conditional SVR projection indicates a temporary increase in 2025 followed by lower and comparatively stable values during 2026–2030 (Fig. 7). The timing of the 2025 increase is approximately 11 years after the 2014 maximum, but the limited number of observed peaks does not permit formal confirmation of a recurrent 10–12-year cycle.

Across all three risk categories, MAE and RMSE provide direct estimates of out-of-time predictive error, whereas the standard deviations and observed ranges in Table 1 characterize variability in the underlying population records. The 2025–2030 trajectories are point projections and do not include formal prediction intervals. Given the near-zero or negative test-period R^2 values, these projections should be interpreted as conditional indications of possible temporal patterns rather than as precise deterministic forecasts.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

4. DISCUSSION

This study used 7192 monitoring records from 25 agroclimatic zones to examine how harmful non-gregarious locust populations respond to environmental variability and whether one forecasting algorithm can be applied uniformly across Kazakhstan. The hypotheses were only partly supported. The trailing three-period population average was the strongest correlate in all risk categories ($r=0.575–0.671$), whereas all individual environmental correlations were weak ($|r|\leq0.104$). Recent population history was therefore more informative than any single climatic variable. Risk categories nevertheless differed in outbreak amplitude, model error, and the environmental variables appearing next in the rankings. Ecological stratification is therefore useful, but the data do not show that temperature alone controls high-risk zones or that moisture alone controls low-risk zones).

The spatially differentiated pattern observed in Kazakhstan agrees with recent work from Central Asia showing that locust responses to climate are geographically non-uniform. Wu et al. (2022) projected spatially heterogeneous changes and an overall contraction of suitable habitats for typical locust species across Kazakhstan and Xinjiang under future climate scenarios. Their habitat-suitability analysis complements, but is not directly comparable with, the continuous population-density forecasts developed in the present study. Klein et al. (2023) showed in Pavlodar Region that meteorological, land-management, and remote-sensing information can be converted into standardized geospatial units for practical survey prioritization. The present zonal framework supports this direction, but its low test-period R^2 values show that climatic and satellite predictors alone were insufficient for reliable density forecasting. Studies from Africa have mainly predicted desert locust presence, breeding habitat, or swarm movement. Sun et al. (2022), Cornejo-Bueno et al. (2023) and Marescot et al. (2025) used lagged temperature, rainfall, soil moisture, and vegetation data for short-term presence prediction. Retkute et al. (2024) integrated breeding-site selection, stage-specific development, feeding, and wind-driven dispersal. These systems address a migratory, phase-transforming species and usually predict presence or movement, whereas the response variable in the present study represents the density of a resident multi-species complex. This more heterogeneous and strongly right-skewed forecasting target helps explain the lower out-of-time explanatory power of the present regression models compared with many classification-based desert locust studies.

Evidence from Europe and other arid ecosystems also supports a combined climate–landscape interpretation. During the recent Moroccan locust outbreak in Sardinia, Italy, drought interacted with increasing areas of uncultivated or abandoned land to create favourable conditions for population increase (Klein et al., 2022). This comparison is important because land use was not included in the present models and may account for part of the unexplained variation. In eastern Australia, Mangeon et al. (2020) found that Australian plague locust abundance increased in productive grasslands but declined rapidly under very hot and dry conditions; their spatial outputs also included uncertainty estimates. Humphreys et al. (2022) likewise reported geographically variable recruitment responses of migratory grasshoppers under projected climate change in the western United States. Together, these studies support region-specific modelling and show that climatic effects are modified by vegetation productivity, land use, population history, and spatial structure.

4.1. Biological Interpretation of the Hydrothermal Ranges

The exploratory land-surface temperature range of 30–35°C is biologically plausible because warm surface and soil conditions can accelerate egg development, hatching, and nymphal growth. However, land-surface temperature is not equivalent to air temperature, insect body temperature, or the temperature inside an egg pod. Because microclimatic conditions vary with soil properties, vegetation cover, topography, solar exposure, and time of day, the identified range should be interpreted as a dataset-specific surface-temperature window rather than as a universal physiological optimum. Recent pest-risk research similarly emphasizes that the temperatures experienced by insects may differ substantially from standard meteorological measurements (Mosedale et al., 2024).

Precipitation of 20–50 mm may represent a balance between the moisture required for egg-pod development and vegetation growth and the adverse effects of hydrological extremes. Rainfall must be sufficient to moisten the oviposition layer and stimulate the emergence of food plants, whereas excessive precipitation may waterlog egg-pod sites, cool the soil, reduce exposed oviposition habitat, and increase mortality. Conversely, very low rainfall may promote egg desiccation and limit food availability after hatching. Comparable intermediate responses have been reported in East African desert locust forecasting and Australian plague locust abundance modelling, although the species, response variables, and spatial scales differ from those of the present study. The normalized soil-moisture range of 0.60–0.75 represents intermediate moisture within the present dataset; it does not correspond to 0.60–0.75 $m^3\ m^{-3}$ or to 60–75% field soil-water content. It therefore cannot be compared directly with raw volumetric soil-moisture thresholds reported for desert locust oviposition. Similarly, SPI values of 0–0.5 represent near-normal to slightly wet conditions rather than a fixed biological breakpoint. The identified LST, precipitation, soil-moisture, and SPI ranges should therefore be treated as descriptive screening indicators for intensified surveillance and not as independently validated treatment thresholds. Field confirmation remains necessary before any phytosanitary intervention is initiated.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

4.2. Model Performance and Forecasting Uncertainty

Model comparison provided only partial support for the hypothesis that nonlinear algorithms would outperform conventional approaches. In the high-risk category, Gradient Boosting produced the lowest RMSE (2.636 individuals m^{-2}) and the highest R^2 value (0.005), whereas SVR produced the lowest MAE (1.307 individuals m^{-2}). The RMSE values of the two algorithms were almost identical. SVR minimized both MAE and RMSE in the moderate- and low-risk categories, but R^2 remained negative (−0.023-0.023−0.023 and −0.010-0.010−0.010, respectively). Model selection is therefore relative: near-zero or negative R^2 values indicate limited test-period variance explanation and, in the moderate- and low-risk categories, no improvement over a mean-only benchmark in terms of R^2.

These findings differ from several desert locust studies that have reported stronger predictive performance for binary occurrence or habitat-suitability classification. This distinction is methodologically important because predicting presence or absence is generally less complex than forecasting the continuous population density of a multi-species locust assemblage characterized by numerous near-zero observations and occasional extreme outbreaks. Predictive performance may be further constrained by uneven monitoring intensity, variation in control interventions, differences in species composition, spatial aggregation of observations, and unmeasured ecological processes. Consequently, lower MAE or RMSE values for a particular algorithm indicate improved relative predictive performance but should not, in isolation, be interpreted as evidence of high operational forecasting accuracy. The strong contribution of the trailing population average indicates substantial short-term temporal persistence in locust abundance. However, this relationship should not necessarily be interpreted as a direct biological mechanism because it may also capture the effects of unmeasured processes, including egg-bank carryover, previous control activities, grazing pressure, vegetation composition, natural enemies, and variation in survey intensity. Similarly, the apparent intervals between major population peaks observed in the annual time series should be regarded as descriptive temporal patterns rather than evidence of established periodic cycles. With only 22 years of observations and relatively few major outbreak peaks, the available time series is insufficient to robustly confirm recurrent 5–7-, 7–10-, or 10–12-year cycles. Longer monitoring records combined with formal spectral, wavelet, and autocorrelation analyses would be required to determine whether such periodicity exists. The projections for 2025–2030 should therefore be interpreted as conditional point projections rather than deterministic outbreak forecasts. They are generated under assumptions regarding the continuation of the modeled relationships and do not currently incorporate formal prediction intervals or probabilistic estimates of forecast uncertainty. Before operational deployment, future versions of the framework should quantify predictive uncertainty using appropriate approaches such as bootstrap resampling, Bayesian inference, quantile regression, or conformal prediction. Model performance should also be benchmarked against simpler alternatives, including persistence and climatological-mean models, as well as hurdle or zero-inflated approaches that may be particularly appropriate for population-density data containing a high proportion of near-zero observations. The spatial abundance modeling approach of Mangeon et al. (2020), in which abundance estimates were accompanied by measures of uncertainty, provides a useful example of how forecast reliability can be communicated transparently to phytosanitary managers and other decision makers. Overall, these limitations do not negate the value of the proposed ML–GIS framework but define its appropriate scope of interpretation. At its current stage, the framework is best considered a risk-oriented decision-support and scenario-projection tool, rather than a deterministic system for predicting the exact magnitude and timing of future locust outbreaks.

4.3. Practical Implications of the ML–GIS Framework

Fig. 8 should be interpreted as a proposed operational architecture rather than as a fully deployed platform. A practical implementation could follow three stages. First, state phytosanitary services would update the database with field density, developmental stage, meteorological observations, and satellite indicators and generate zone-level estimates of P_{abs} and risk. Second, these outputs would be used to prioritize survey routes, particularly where predicted population density approaches an official action threshold or where recent field information is incomplete. Third, control measures would be considered only after ground teams confirmed species composition, density, developmental stage, occupied area, and the exposure of crops or pastures. For farmers, the system should provide simple map-based alerts showing the risk category, affected administrative unit, date of the latest update, and recommended scouting action. It should not automatically prescribe pesticides, application rates, or product selection. For phytosanitary agencies, the interface should retain continuous P_{abs} estimates, recent observations, the selected model, data-quality flags, and, once developed, uncertainty intervals. International operational experience indicates that remote sensing is most useful when it reduces and prioritizes the area requiring ground surveys rather than replacing field observation (Klein et al., 2023). High-risk zones, particularly zones (23), (42), and (75), may receive priority when personnel and monitoring resources are limited. However, the previously proposed bi-weekly survey frequency and the claimed 35–40% cost saving were removed because neither was evaluated in this study. Monitoring frequency should instead be determined by national phytosanitary protocols, species phenology, current risk, data

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

availability, and survey capacity. A feedback mechanism should also be incorporated so that verified field observations are returned to the database and used for periodic recalibration of the models.

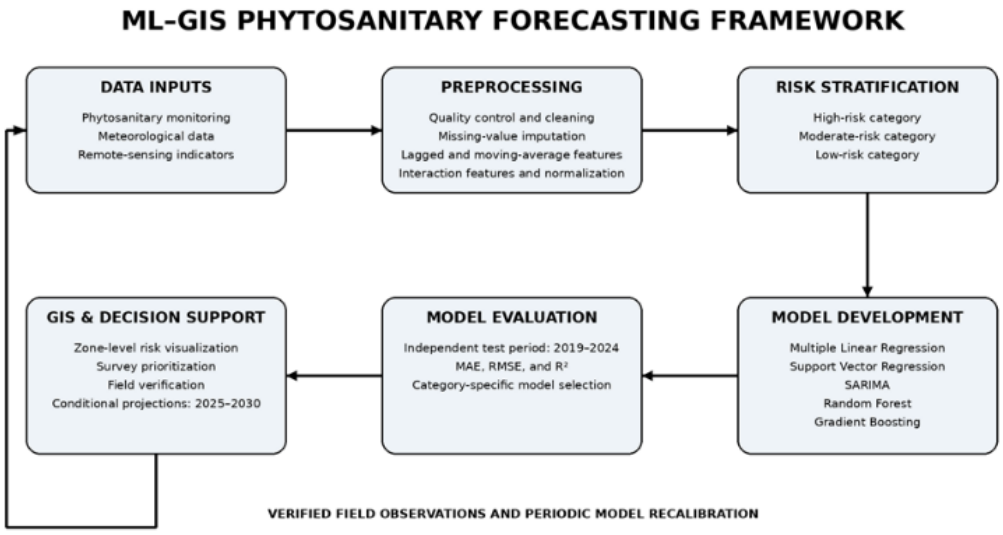

Fig. 8: Conceptual architecture of the proposed ML–GIS phytosanitary forecasting framework. The system integrates monitoring, climatic, and remote-sensing data, preprocessing, risk stratification, machine-learning and statistical modelling, model evaluation, GIS-based decision support, field verification, and periodic model recalibration.

4.4. Limitations and Research Directions

The principal limitations of this study are the aggregation and uneven distribution of historical records, the extreme right-skewness of population density, the use of a combined multi-species response, the absence of formal prediction intervals, and the low or negative test-period R^2 values. Risk groups were defined from long-term mean population density, and no independent geographic validation was performed outside Kazakhstan. Land use, grazing pressure, control history, crop distribution, soil texture, natural enemies, and species-specific phenology were not incorporated. The limited temporal coverage also restricts inference about low-frequency cycles, while the continuation scenario used for 2025–2030 does not constitute a climate-change projection. Future research should combine hierarchical spatiotemporal models with species- and life-stage-specific data, explicit land-use and management variables, microclimatic estimates, and uncertainty-aware forecasting. External validation in neighbouring Central Asian countries would test the transferability of the risk stratification and model rankings. Annual recalibration using newly verified monitoring records would further determine whether forecasting skill improves over time. The present ML–GIS framework should therefore be regarded as a structured decision-support prototype: it identifies useful spatial strata and candidate hydrothermal warning ranges, but field verification and further validation remain essential before operational control decisions are automated.

5. CONCLUSION

This study developed and evaluated an ecologically stratified ML–GIS framework for forecasting harmful non-gregarious locust populations across Kazakhstan. The analysis was based on 7,192 phytosanitary monitoring records collected in 25 agroclimatic zones during 2003–2024. The trailing three-period moving average of population density was the strongest correlate of current population density in all three risk categories ($r=0.575–0.671$), whereas the correlations with individual environmental predictors were weak ($|r|\leq0.104$). These findings indicate that recent population history was more informative than any single climatic variable and that environmental effects are likely to operate through nonlinear interactions and temporal lags. Model performance differed among risk categories. In the high-risk category, Gradient Boosting produced the lowest RMSE and the highest R^2, whereas SVR produced the lowest MAE. SVR minimized both MAE and RMSE in the moderate- and low-risk categories. However, the near-zero or negative test-period R^2 values demonstrate that these models had only limited out-of-time explanatory capacity. Their selection therefore reflects relative performance among the tested approaches rather than high forecasting accuracy. Land-surface temperatures of 30–35°C, precipitation totals of 20–50 mm, normalized soil-moisture values of 0.60–0.75, and SPI values of 0–0.5 were associated with comparatively elevated population densities. These ranges should be interpreted as dataset-specific surveillance indicators rather than universal physiological or treatment thresholds. The proposed framework may assist phytosanitary agencies in prioritizing field surveys and allocating monitoring resources among agroclimatic zones. Nevertheless, the 2025–2030 outputs are conditional point projections without formal prediction intervals and should not replace field verification. Further development requires species- and life-stage-specific observations, land-use and management variables, external geographic validation, explicit uncertainty estimation, and periodic recalibration using newly verified monitoring data.

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

Declarations

Funding: The research was carried out with financial support from the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan for 2024–2026 within the framework of the scientific project IRN AP22784639 "Development of phytosanitary forecasting models for harmful grasshoppers' population in Kazakhstan based on machine learning algorithms and GIS technologies".

Acknowledgement: The authors would like to express their sincere gratitude to the Editor and Reviewers of Agrobiological Records for their valuable comments, constructive suggestions, and careful evaluation, which contributed to improving the quality and scientific presentation of this manuscript.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Data Availability: Data will be available upon request.

Ethics Statement: This study was based on historical phytosanitary monitoring records, meteorological observations, remote-sensing products, and GIS data. No research involving human participants, vertebrate animals, or experimental manipulation of live animals was conducted.

Author's Contributions: Conceptualization, K.B. and A.R.; methodology, K.B., A.R., and Z.A.; investigation and phytosanitary data collection, K.B., A.B., Z.A., and A.Z.; data curation, A.B., Z.A., and A.Z.; software, statistical analysis, and machine-learning modelling, A.R. and A.D.; GIS analysis and visualization, A.R., A.Z., and A.D.; writing – original draft preparation, K.B. and A.R.; writing – review and editing, A.B., Z.A., A.Z., and A.D.; supervision and project administration, K.B. All authors have read and approved the final version of the manuscript.

Generative AI Statements: The authors declare that no Gen AI/DeepSeek was used in the writing/creation of this manuscript.

Publisher's Note: All claims stated in this article are exclusively those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated/assessed in this article or claimed by its manufacturer is not guaranteed or endorsed by the publisher/editors.

REFERENCES

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. Trends in Ecology & Evolution, 22(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Azhbenov, V. K., Baibussenov, K. S., Sarbaev, A. T., & Harizanova, V. B. (2015). Preventive approach of phytosanitary control of locust pests in Kazakhstan and adjacent areas. In Proceedings of the International Conference on Agricultural, Ecological and Medical Sciences (AEMS-2015) (pp. 33–37). https://doi.org/10.15242/IICBE.C0215122

Baibussenov, K. S., & Amanbay, Z. Z. (2025). Improvement of methods of phytosanitary control of locusts in the moderately humid moderately warm agro-climatic zone of Kazakhstan based on the use of geoinformation and information technologies. Eurasian Agrotechnical Journal, 2/1(126), 17–28. https://doi.org/10.51452/kazatu.2025.2/1(126).1912

Baibussenov, K., Amanbay, Z., Bekbaeva, A., Azhbenov, V., & Rustembayev, A. (2025). Applying machine learning models to predict population dynamics of harmful non-gregarious locusts across Kazakhstan agroclimatic zones. World Journal of Environmental Biosciences, 14(4), 100–111. https://doi.org/10.51847/5oaisyhbcp

Baibussenov, K., Bekbayeva, A., & Azhbenov, V. (2022). GIS-based potential distribution modeling for harmful non-gregarious locusts in agricultural areas of Northern Kazakhstan to improve preventive pest management. In Proceedings of the 24th Paris International Conference on Advances on Chemical, Agricultural, Biological & Environmental Sciences (CABES-22) (pp. 48–56). https://doi.org/10.17758/EARES11.EAP0422106

Baibussenov, K., Bekbayeva, A., & Azhbenov, V. (2024). Simulation of favourable habitats and potential spread for non-gregarious locust pests for preventive measures in North Kazakhstan based on satellite data. Innovations in Biological Science, 6, 153–174. https://doi.org/10.9734/bpi/ibs/v6/715

Congalton, R. G., & Green, K. (2019). Assessing the accuracy of remotely sensed data: Principles and practices. CRC Press.

Cornejo-Bueno, L., Pérez-Aracil, J., Casanova-Mateo, C., Sanz-Justo, J., & Salcedo-Sanz, S. (2023). Machine learning classification–regression schemes for desert locust presence prediction in Western Africa. Applied Sciences, 13(14), 8266. https://doi.org/10.3390/app13148266

de Oliveira Aparecido, L. E., de Souza Rolim, G., da Silva Cabral de Moraes, J. R., Costa, C. T. S., & de Souza, P. S. (2020). Machine learning algorithms for forecasting the incidence of pests and diseases in Coffea arabica. International Journal of Biometeorology, 64(4), 671–688. https://doi.org/10.1007/s00484-019-01856-1

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

Didan, K., Munoz, A. B., Solano, R., & Huete, A. (2019). MODIS vegetation index user's guide (MOD13 series). Vegetation Index and Phenology Lab, University of Arizona. https://vip.arizona.edu/MODIS_ATBD.php

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37(12), 4302–4315. https://doi.org/10.1002/joc.5086

Gahramanov, S., Ismayilova, V., Taghiyev, M., Mammadov, H., & Musayeva, K. (2024). Study of the resistance of economically important pine species to dangerous pests on the Absheron Peninsula. BIO Web of Conferences, 100, 04035. https://doi.org/10.1051/bioconf/2024100

Gómez, D., Salvador, P., Sanz, J., Casanova, C., Taratiel, D., & Casanova, J. L. (2018). Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture. Journal of Applied Remote Sensing, 12(3), 036011. https://doi.org/10.1117/1.JRS.12.036011

Goodchild, M. F. (2009). Geographic information systems and science: Today and tomorrow. Annals of GIS, 15(1), 3–9. https://doi.org/10.1080/19475680903250715

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/b94608

Humphreys, J. M., Srygley, R. B., & Branson, D. H. (2022). Geographic variation in migratory grasshopper recruitment under projected climate change. Geographies, 2(1), 12–30. https://doi.org/10.3390/geographies2010003

Ibrahim, E. A., Salifu, D., Mwalili, S., Dubois, T., Collins, R., & Tonnang, H. E. (2022). An expert system for insect pest population dynamics prediction. Computers and Electronics in Agriculture, 198, 107124. https://doi.org/10.1016/j.compag.2022.107124

Klein, I., Cocco, A., Uereyen, S., Mannu, R., Floris, I., Oppelt, N., & Kuenzer, C. (2022). Outbreak of Moroccan locust in Sardinia (Italy): A remote sensing perspective. Remote Sensing, 14(23), 6050. https://doi.org/10.3390/rs14236050

Klein, I., Oppelt, N., & Kuenzer, C. (2021). Application of remote sensing data for locust research and management—A review. Insects, 12(3), 233. https://doi.org/10.3390/insects12030233

Klein, I., Uereyen, S., Eisfelder, C., Pankov, V., Oppelt, N., & Kuenzer, C. (2023). Application of geospatial and remote sensing data to support locust management. International Journal of Applied Earth Observation and Geoinformation, 117, 103212. https://doi.org/10.1016/j.jag.2023.103212

Lawton, D., Scarth, P., Deveson, E., Piou, C., Spessa, A., Waters, C., & Cease, A. J. (2022). Seeing the locust in the swarm: Accounting for spatiotemporal hierarchy improves ecological models of insect populations. Ecography, 2022(2), e05763. https://doi.org/10.1111/ecog.05763

Mallie, D. T., Demissie, S. T., Beza, S. A., Erena, S. H., & Dejene, S. W. (2026). Dynamics of vegetation responses to desert locust (Schistocerca gregaria) plagues of 2019–2021 in Wabe Shebelle River Basin of Ethiopia. Scientific African, 33, e03491. https://doi.org/10.1016/j.sciaf.2026.e03491

Mamo, D. K., Kinyanjui, M. N., & Siewe, N. (2025). Modeling desert locust population dynamics: A climate driven approach using machine learning. Ecological Modelling, 511, 111363. https://doi.org/10.1016/j.ecolmodel.2025.111363

Mangeon, S., Spessa, A., Deveson, E., Darnell, R., & Kriticos, D. J. (2020). Daily mapping of Australian plague locust abundance. Scientific Reports, 10, 16915. https://doi.org/10.1038/s41598-020-73897-1

Marescot, L., Fernandez, E., Dridi, H., Benahi, A. S., Hamouny, M. L., Maeno, K. O., Escorihuela, M., Paolini, G., & Piou, C. (2025). A forecasting model for desert locust presence during recession period, using real-time satellite imagery. Remote Sensing Applications: Society and Environment, 37, 101497. https://doi.org/10.1016/j.rsase.2025.101497

Mingalev, D. E. (2021). Agroclimatic zoning of Russia and Kazakhstan under current climate change. Geography and Natural Resources, 42(2), 24–31. https://doi.org/10.15372/GIPR20210203

Mosedale, J. R., Eyre, D., Korycinska, A., Everatt, M., Grant, S., Trew, B., Kaye, N., Hemming, D., & Maclean, I. M. D. (2024). Mechanistic microclimate models and plant pest risk modelling. Journal of Pest Science, 97, 1749–1766. https://doi.org/10.1007/s10340-024-01777-y

Peters, D. P. C., Bestelmeyer, B. T., Knapp, A. K., Herrick, J. E., Monger, H. C., & Havstad, K. M. (2009). Approaches to predicting broad-scale regime shifts using changing pattern-process relationships across scales. In S. Miao, S. Carstenn, & M. Nungesser (Eds.), Real world ecology: Large-scale and long-term case studies and methods (pp. 47–72). Springer. https://doi.org/10.1007/978-0-387-77942-3_3

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Retkute, R., Thurston, W., Cressman, K., & Gilligan, C. A. (2024). A framework for modelling desert locust population dynamics and large-scale dispersal. PLOS Computational Biology, 20(12), e1012562. https://doi.org/10.1371/journal.pcbi.1012562

Singh, A. K., Yeasin, M., Paul, R. K., Paul, A. K., & Sarkar, A. (2024). Dynamic ensemble-based machine learning models for predicting pest populations. Frontiers in Applied Mathematics and Statistics, 10, 1435517. https://doi.org/10.3389/fams.2024.1435517

Sun, R., Huang, W., Dong, Y., Zhao, L., Zhang, B., Ma, H., Geng, Y., Ruan, C., Xing, N., Chen, X., & Li, X. (2022). Dynamic forecast of desert locust presence using machine learning with a multivariate time lag sliding window technique. Remote Sensing, 14(3), 747. https://doi.org/10.3390/rs14030747

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074

ISSN: 2708-7182 (Print); ISSN: 2708-7190 (Online)
Open Access Journal

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD—A platform for ensemble forecasting of species distributions. Ecography, 32(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Wu, R., Guan, J., Wu, J., Ju, X., An, Q., & Zheng, J. (2022). Predictions based on different climate change scenarios: The habitat of typical locust species is shrinking in Kazakhstan and Xinjiang, China. Insects, 13(10), 942. https://doi.org/10.3390/insects13100942

Xiao, Q., Li, W., Kai, Y., Chen, P., Zhang, J., & Wang, B. (2019). Occurrence prediction of pests and diseases in cotton on the basis of weather factors by long short term memory network. BMC Bioinformatics, 20(Suppl 25), 688. https://doi.org/10.1186/s12859-019-3262-y

Citation: Baibussenov, K., Rustembayev, A., Bekbayeva, A., Amanbay, Z., Zhubatkanov, A., & Dzhumagulov, A. (2026). Agroclimatic variability and machine Learning–GIS forecasting of harmful Non-gregarious locusts in Kazakhstan. Agrobiological Records, 25, 413-432. https://doi.org/10.47278/journal.abr/2026.074