STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

Statistical and ANN-Based Modeling for Desertification Risk Prediction in Semi-Arid Regions: A Case Study of Nineveh Governorate

Ziadoon Mohand Khaleel[1,*], Safa Jawad Abed[2], Jalal Abdulkareem Sultan[2], Noor Marwan Ahmeed[1]

[1] *Department of Petroleum Reservoir Eng.,College of Petroleum and Mining Engineering, University of Mosul,Mosul,Iraq*
[2] *Nineveh Agriculture Directorate, Iraq*

Abstract Desertification and land degradation threaten food, water, and livelihood security across Iraq's semi-arid north. We develop a station-based, hybrid framework for operational desertification-risk assessment in Nineveh Governorate using multi-decadal observations from Mosul, Tal Afar, and Rabiya (1991–2024). The approach couples distribution-free trend diagnostics Mann Kendall with Sen's slope and persistence control via trend-free pre-whitening and effective-sample-size corrections with an interpretable composite Desertification Risk Index (DRI) built from directionally normalized indicators (temperature, humidity, wind, sunshine, rainfall) and objective CRITIC (Criteria Importance Through Intercriteria Correlation) weights. For prediction, we generate horizon-explicit DRI forecasts at 1, 3, 5, and 10 years using leakage-free rolling-origin splits with training-only transformations and chronological refitting. Baselines (Persistence, Climatology) are compared with ARIMA, artificial neural networks (ANN), Random Forest (RF), and XGBoost using $\mathrm{RMSE}/\mathrm{MAE}/R^2$ and Skill relative to Persistence; pairwise differences are assessed with Diebold–Mariano tests. Skill is non-negative across all 12 station-by-horizon cases; RF dominates at medium long horizons especially in Tal Afar and Rabiya with Skill reaching ≈ 0.35–0.38 at Tal Afar (3–5 years; RF/XGB) and ≈ 0.39 at Rabiya (10 years; XGB) , while ANN is competitive at short leads (e.g., Tal Afar 1-year; Mosul 3-year); ARIMA is only competitive at Mosul (1-year). DM tests indicate significant improvements in a subset of cases, while many differences are not significant under short annual samples. The framework yields reproducible diagnostics and horizon-aware outlooks that can augment persistence/climatology in early-warning workflows; priority extensions include integrating remote-sensing predictors, homogenization, seasonal targets, and probabilistic verification.

Keywords Artificial neural networks, CRITIC weighting, Desertification Risk Index, Mann–Kendall trend test, Semi-arid regions

AMS 2010 subject classifications 62H30, 62M20

DOI: 10.19139/soic-2310-5070-2985

1. Introduction

Desertification and land degradation are among the most consequential environmental risks of the twenty first century, disproportionately affecting drylands where water scarcity, heat extremes, and fragile agro ecosystems intersect. The Intergovernmental Panel on Climate Change reports heightened exposure and vulnerability across arid and semi-arid regions as warming and drying compound human pressures on land systems and threaten food, water, and livelihood security [1]. In the Eastern Mediterranean and Middle East (including Iraq), multiple lines of evidence indicate rapid warming, intensifying heatwaves, and an escalation of compound hot–dry extremes that accelerate degradation processes [2]. Within Iraq, the northern governorates, including Nineveh, are agriculturally and strategically significant yet increasingly climate-sensitive; recent assessments document rising temperatures,

*Correspondence to: ziadoon Mohand Khaleel (Email: ziadoon.khaleel@uomosul.edu.iq). Department of Petroleum Reservoir Eng.,College of Petroleum and Mining Engineering, University of Mosul,Mosul,Iraq

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

seasonal drying, and shifts in rainfall regimes together with discontinuities in some station records that complicate trend detection and risk monitoring [3].

Against this backdrop, operational desertification-risk assessment requires (i) diagnostics that are robust to non-Gaussian behavior and autocorrelation in hydro climate series, and (ii) prediction systems that integrate multiple indicators and deliver actionable lead times. Non-parametric trend tests such as Mann–Kendall with Sen's slope have become standard tools for monotonic change detection in hydro-climate time series[4, 5] , with modified formulations recommended under persistence and scaling [6]. Composite indices provide a transparent way to synthesize temperature, humidity, wind, sunshine, and rainfall into a single diagnostic; objective weighting schemes such as CRITIC (with entropy as a robustness check) reduce subjectivity by exploiting contrast intensity and inter-criterion conflict directly from the data matrix [7, 8]. In parallel, artificial neural networks and related learners have shown promise in emulating nonlinear climate–land relationships when trained on multi-source predictors [9]-[11].

Building on these strands, this study develops a station-based, hybrid framework for desertification-risk assessment in Nineveh Governorate. We (1) curate multi-decadal meteorological observations from Mosul, Tal Afar, and Rabiya; (2) perform distribution-free trend diagnosis on constituent indicators; (3) construct a Desertification Risk Index (DRI) via directional normalization and objective weights (CRITIC, with entropy sensitivity); and (4) evaluate a time-aware prediction component for multi-year outlooks using chronological splits and horizon-explicit verification. The framework aims to balance interpretability (trend tests, transparent weights) with predictive utility and reproducibility for semi-arid risk monitoring in Iraq and comparable settings. To ensure statistical rigor and sound algorithmic analysis, forecasts are benchmarked against a persistence baseline and differences in predictive accuracy are tested using Diebold–Mariano; model interpretability diagnostics (e.g., SHAP) are reported, and we outline a probabilistic extension evaluated (e.g., CRPS) [12, 14] .

This work is motivated by the need for transparent, statistically rigorous, and actionable risk information for semi-arid planning in Nineveh. Our contributions are presented within a single, reproducible pipeline: (i) an interpretable composite DRI with objective CRITIC weights and entropy-based sensitivity; (ii) distribution-free trend inference (Mann–Kendall, Sen's slope) with adjustments for serial dependence via trend-free pre-whitening; (iii) chronological, leakage-free evaluation via rolling/forward-origin splits and horizon-explicit verification for 1–10-year outlooks; (iv) formal forecast benchmarking against a persistence baseline with Diebold–Mariano tests; (v) model interpretability using SHAP to attribute indicator-level effects; and (vi) probabilistic assessment using proper scoring rules (e.g., CRPS) to quantify uncertainty in risk outlooks.

Section (2) reviews trend diagnostics, composite indices, and learning-based risk models; Section (3) details the study area, data, and preprocessing; Section (4) presents methods; Section (5) reports results; Section (6) discusses implications and limitations; Section (7) concludes.

2. RELATED WORK

2.1. Desertification risk diagnostics in drylands

Recent syntheses emphasize that operational desertification risk assessment must reconcile two needs: (i) distribution free trend diagnostics that remain valid under non-Gaussianity and autocorrelation and (ii) multi-indicator risk synthesis with transparent weighting and traceable uncertainty. In practice, Mann–Kendall (MK) and Sen's slope remain the workhorses for monotonic trends in hydro-climate time series, but prewhitening variants are required when persistence is present to control Type-I error and slope bias [19] – [21] . Simulation and method papers show that naïve prewhitening can distort inference, while newer schemes (e.g., DPWMT) improve power and bias across a broad AR(1) range; agency studies also benchmark MK against GLS-based alternatives in autocorrelated series [19, 20]. These results support our choice of MK with trend-free prewhitening (and sensitivity checks), aligning with best-practice recommendations for statistical rigor.

2.2. Composite indices and objective weighting

Composite indices remain central for integrating temperature, humidity, wind, sunshine, and rainfall into a single diagnostic of land-degradation pressure. To reduce subjectivity, objective weighting has advanced beyond

the classic CRITIC method (variance × inter-criterion conflict) to improved formulations (e.g., CRITID) and systematic comparisons across objective schemes [15, 16].
Parallel work on entropy weights clarifies how data normalization and scaling choices materially affect the inferred weights, and offers robust, reproducible pipelines for index construction [17, 18]. We therefore adopt CRITIC as the primary scheme with entropy-based sensitivity analysis, ensuring interpretability and reproducibility.

2.3. Satellite indices and data products relevant to desertification risk

Remotely sensed vegetation condition remains a leading proxy for monitoring stress. The Vegetation Health Index (VHI) a fusion of vegetation (VCI) and thermal (TCI) signals has matured with NOAA's Blended-VHP (AVHRR 1981–2012 + VIIRS 2013–present), supporting weekly global coverage and consistent time series used in drought and crop-stress applications [25]–[27]. In parallel, NDVI time series continue to improve: recent datasets deliver spatiotemporally consistent long-records (e.g., PKU-GIMMS NDVI 1982–2022) and higher-resolution gap-filled products for the MODIS era, enabling stronger links between vegetation dynamics and climate drivers [28, 29] . These developments justify our use of station-based meteorology augmented by VHI/NDVI in sensitivity analyses and validation.

2.4. Machine learning for drought/desertification prediction

A rapidly growing body of work applies machine learning (ML) for drought early warning and land-degradation risk, from tree ensembles and gradient boosting to deep learning. Recent reviews synthesize progress and pitfalls, underscoring the need for chronological validation, baseline benchmarking, uncertainty quantification, and interpretability [11, 30, 35]. In the Middle East and Iraq, case studies span susceptibility mapping with (GRACE/GLDAS + ML, LSTM-based) drought index prediction, and MEDALUS-style sensitivity mapping in northern governorates evidence that hybrid, data driven approaches are gaining traction in climatically stressed, data limited settings [10],[32]–[34]. Region-wide work has also integrated ML with climate-model outputs to improve monitoring and spatial adaptation of dryness indices [31]. These strands motivate our hybrid station based design: (objective-weight DRI + classical/ML) predictors, with careful temporal validation and interpretability.

2.5. Forecast evaluation, uncertainty, and interpretability

For out-of-sample evaluation in time series, the standard is rolling-origin (walk-forward) assessment with refitting, which provides horizon-explicit error distributions and avoids leakage; contemporary guides warn against random K-fold CV and stress baseline comparisons [22]–[24].To compare models formally, the Diebold–Mariano (DM) framework ideally with the Harvey–Leybourne–Newbold small sample correction tests equality of predictive accuracy under serial correlation and overlapping errors [22, 37] . Beyond point forecasts, proper scoring rules such as the Continuous Ranked Probability Score (CRPS) provide coherent evaluation of predictive distributions; modern toolkits (e.g., scoringRules) and recent reviews document implementation and distribution-specific formulae for CRPS and related scores [38]. Finally, model interpretability (e.g., SHAP) has become a recommended complement to skill-only reporting in environmental ML, clarifying indicator-level contributions and supporting decision transparency; we therefore pair skill/DM/CRPS with SHAP summaries.

2.6. Regional context: Eastern Mediterranean & Iraq

The Eastern Mediterranean and Middle East (EMME) is a recognized climate change hotspot with intensifying compound hot dry extremes [1, 2]. Within Iraq, remote sensing and multi-index assessments indicate spatial heterogeneity and rising exposure across agricultural governorates, including Nineveh [3, 24, 25, 33]. Prior studies typically emphasize a single strand trend diagnostics, composite indices, or prediction whereas fewer integrate all three under leakage free, horizon-explicit evaluation with strong baselines [7, 8, 11, 21, 30]. The present framework addresses this gap by combining distribution-free trend tests with objective multi-indicator weighting (CRITIC, with entropy sensitivity) and chronological forecasting at multiple lead times [4]-[6],[7, 8, 16],[22]-[24] .

3. STUDY AREA AND DATA

3.1. Geographical setting

The study focuses on Nineveh Governorate in northern Iraq, with administrative seat Mosul and an area of approximately 37,323 km^2, bordering Syria to the northwest and the governorates of Duhok, Erbil, Kirkuk, Salah al-Din, and Al-Anbar [39]. Climatically, Nineveh is predominantly semi-arid steppe (BS) under the Köppen–Geiger classification, with hot, dry summers and cool, wetter winters [40]. Topography is characterized by the Tigris valley lowlands around Mosul and gently rising terrain toward the Syrian border; elevations used for mapping and context were extracted from the SRTM digital elevation model [41]. All geospatial layers are projected to WGS84 (EPSG:4326) and clipped to the administrative boundary of Nineveh [39].

3.2. Observation network and variables

We analyze multi-decadal station observations from three synoptic locations Mosul, Tal Afar, and Rabiya operated within Iraq's national meteorological framework. The baseline variable set comprises daily to monthly maximum/minimum air temperature (°C), relative humidity (%), wind speed (m s^{-1}), bright sunshine hours (h), and rainfall (mm). Data access and coverage windows per station are coordinated via Iraqi national services and respected chronologically in all analyses [50]. Across the three stations, the annual records exhibit identical temporal coverage (1991–2024; N = 34 years). This uniformity justifies applying a single, station wise preprocessing, orientation/normalization, and evaluation protocol across sites and removes the need for a dedicated coverage table.

3.3. Auxiliary gridded/reanalysis and remote-sensing products

To cross-validate station series, fill small gaps where permissible, and provide spatial context, we incorporate: ERA5-Land for near surface meteorology and land-surface fields [42]; CHIRPS for precipitation [43]; CRU TS v4 for long records of temperature/precipitation and vapour pressure [45]; and MODIS vegetation indices (NDVI/EVI) for vegetation dynamics [44]. Where evapotranspiration or root-zone soil moisture are referenced for context, we use GLEAM [46]. Unless otherwise noted, ERA5-Land (≈0.1°), CHIRPS v2.0 (0.05°), CRU TS v4.08 (0.5°), MODIS MOD13A2 v6.1 (1 km/16-day), and GLEAM v3.x (0.25°) are used only for validation/diagnostic mapping and not ingested into forecast models, thereby avoiding information leakage (Sections 5–6). All external datasets are reprojected to WGS84 and subset to the Nineveh administrative boundary [39].

3.4. Indicators, units, and risk orientation

For the Desertification Risk Index (DRI), station level indicators are directionally normalized so that larger values consistently indicate higher risk: increases in $T_{\max}$, $T_{\min}$, wind speed, and sunshine raise risk, whereas higher rainfall reduces risk; lower relative humidity indicates increased dryness and higher risk. Variables are standardized at the monthly scale relative to local climatology prior to objective weighting (Section 4). Units and abbreviations follow WMO conventions [47].

3.5. Quality control, homogeneity, and missing data

Quality control follows the WMO Guide to Climatological Practices (screening for physically implausible values, internal-consistency checks, and neighbor checks where feasible [47]. Potential inhomogeneities due to metadata gaps, relocations, or instrument changes are screened using the non-parametric Pettitt test, corroborated with the SNHT for mean shifts; flagged breaks are documented, and adjustments are applied only when supported by metadata/evidence [47]-[49]. Short gaps ($\leq$ 3 consecutive months) are handled with conservative, within station procedures (ratio or seasonal-mean substitution) that never cross train/test boundaries (Sections 5–6).

3.6. Temporal aggregation and splits

Daily observations are aggregated to monthly values sums for rainfall and means for other variables and, where specified, further aggregated to annual indices for descriptive analyses. Model evaluation uses strictly chronological splits with a hold out period reserved for out of sample testing; auxiliary datasets in subsection (3.3) are used only for cross checks and descriptive maps and are not allowed to leak information into forecasts.

4. METHODS

4.1. Pre-processing and temporal aggregation

Monthly observations at Mosul, Tal Afar, and Rabiya were screened for physically impossible values and keying errors and then aggregated to the annual scale for descriptive analyses and forecasting. Rainfall and sunshine were summed; air temperatures, relative humidity, and wind speed were averaged consistent with WMO reporting practice [47]. All processing was performed station-wise to avoid any cross site leakage. All transformations used in modeling (normalization, weighting, feature construction) were computed on training folds only in every split.

4.2. Trend diagnostics and persistence control

Monotonic trends in each indicator (and in the composite index) were assessed using the Mann–Kendall (MK) test with Sen's slope as a robust rate estimator [4, 5]. Because hydro climate series often exhibit serial correlation, we applied persistence aware adjustments prior to inference: (i) Trend Free Pre Whitening (TFPW) detrend, AR(1) pre-whiten, re-add trend followed by MK on the adjusted series [6]; and (ii) an effective-sample-size (ESS) variance inflation for MK, parameterized by the AR(1) estimate, as recommended in simulation/agency studies for autocorrelated series [20]. Unless noted otherwise, Sen's slope is computed on the original series, while MK p-values are persistence adjusted. Sensitivity to the pre-whitening scheme was checked against improved variants in the literature [19].

4.3. Indicator orientation and normalization

To impose consistent "higher-is-higher-risk" semantics, each station level indicator was directionally normalized on the training data. For risk-increasing variables (e.g., $T_{\max}$, $T_{\min}$, wind, sunshine), we used a min–max transform; for risk reducing variables (rainfall, relative humidity), we inverted the scale so larger values still indicate higher risk (see Eq.1):

$$z_{j,t} = \begin{cases} \dfrac{x_{j,t} - \min_{\text{train}} x_j}{\max_{\text{train}} x_j - \min_{\text{train}} x_j}, & \text{(risk-increasing)}, \\ 1 - \dfrac{x_{j,t} - \min_{\text{train}} x_j}{\max_{\text{train}} x_j - \min_{\text{train}} x_j}, & \text{(risk-reducing)}. \end{cases} \tag{1}$$

Normalization bounds were estimated on training folds only at each origin to preclude information leakage.

4.4. Objective weighting and DRI construction

Weights were derived via CRITIC (objective weighting), which combines each variable's contrast (σ_j) with its inter criterion conflict (low average correlation with others) [7]. Weights are normalized to sum to one (see Eq.2):

$$w_j = \frac{\sigma_j(1 - \bar{\rho}_j)}{\sum_k \sigma_k(1 - \bar{\rho}_k)}. \tag{2}$$

The Desertification Risk Index (DRI) at year tis the CRITIC-weighted sum of oriented indicators (see Eq.3):

$$\text{DRI}_t = \sum_j w_j \, z_{j,t}. \tag{3}$$

As a robustness check, entropy weights were computed but not used for the mainline results; their impact is summarized in sensitivity analyses [8].

4.5. Orecasting targets, features, and splits

We forecast DRI_{t+h} at horizons $h \in \{1, 3, 5, 10\}$ using only information available at time t. Features are lagged, oriented indicators (lags 0–3) constructed after training only normalization and CRITIC weight estimation to prevent information leakage. The choice of lags ≤ 3 is supported by partial autocorrelation and mutual information diagnostics computed on the training folds and found to be consistent across stations. Evaluation follows a leakage free rolling origin (walk-forward) protocol with an initial 15-year training window, a 5-year block of test bases per split, and a 5-year step between origins. At each origin, the entire pipeline (orientation/normalization, objective weighting, feature construction, and model fitting) is refit using the training fold only, and forecasts are issued for $t + h$. Comparators include Persistence $\hat{y}_{t+h} = y_t$, Climatology (training-mean DRI) ARIMA (orders selected by AIC on the training DRI and refit per origin), a feed-forward ANN (e.g., 64–32 units with early stopping), Random Forest, and XGBoost all trained with fixed seeds and light regularization. Metrics are computed on the pooled out of sample predictions per station × horizon in Section (5).

4.6. Performance metrics and skill

For each station × horizon, predictions from all origins are pooled and summarized with RMSE, MAE, and R^2. Skill is reported relative to Persistence (see Eq.4):

$$\text{Skill} = 1 - \frac{\text{RMSE}_{\text{model}}}{\text{RMSE}_{\text{Persistence}}}\,. \tag{4}$$

so positive values indicate improvement over the Persistence baseline. We further report station-wise distributions of per-origin errors to visualize dispersion and potential non-Gaussianity [22]-[24].

4.7. Statistical comparison of models

To assess whether differences vs Persistence are statistically meaningful, we apply the Diebold–Mariano (DM) test with squared-error loss and Newey–West variance; the truncation lag is $(q = h - 1)$ to reflect the forecast horizon [13]. Given short annual records, we report two-sided p-values and complement DM with the Harvey–Leybourne–Newbold small-sample adjustment where noted [37]. This guards against over interpreting small differences in limited samples [22]-[24].

4.8. Reproducibility

All steps training-only normalization, CRITIC weighting, TFPW/ESS trend diagnostics, rolling-origin splits, metrics, and DM tests are automated in a scripted workflow with fixed random seeds for ML models. Configuration files capture dataset paths, feature definitions, search grids, and evaluation settings, ensuring exact reproducibility.

4.9. Interpretability and sensitivity

We compute global SHAP importance for RF/ANN and local SHAP for selected years to attribute indicator level effects [12]. A one out ablation removes each indicator in turn to quantify its marginal contribution to DRI and to forecast skill at each horizon.

4.10. Lag and architecture selection

Lag sufficiency is supported by PACF and mutual information up to lag 3 for most stations. ANN hyper parameters are selected via chronological random/Bayesian search over the ranges given in SubSection (4.5), monitored by early stopping to mitigate overfitting.

4.11. Uncertainty quantification

We quantify uncertainty in skill using a moving-block bootstrap across origins (block length matched to h), reporting 95% confidence intervals for RMSE/MAE skill. This supplements DM tests and reflects dependence in pooled errors [22]-[24].

4.12. Probabilistic forecasts

To provide decision-useful uncertainty summaries, we fit Quantile RF and Quantile loss MLP to produce $10^{th}/50^{th}/90^{th}$ percentiles. Probabilistic performance is evaluated with the Continuous Ranked Probability Score (CRPS) and reliability/coverage diagnostics [14, 38].

5. RESULTS

5.1. Overview

This section reports the end-to-end evaluation under the rolling-origin (walk-forward) protocol described in Section (4). Figure (1) summarizes the workflow from quality control and monthly to annual aggregation through orientation/normalization, CRITIC weighting, DRI computation, and evaluation implemented with training only transformations and chronological refitting at each origin. Performance is summarized with RMSE, MAE, and R^2 and reported as Skill relative to the Persistence baseline (Eq. 4), following standard practice in forecast verification [22]-[24]. As context, Figure (2) shows station wise DRI time series for Mosul, Tal Afar, and Rabiya over 1991–2024; captions report percent change in DRI relative to a baseline period (1991–2000).

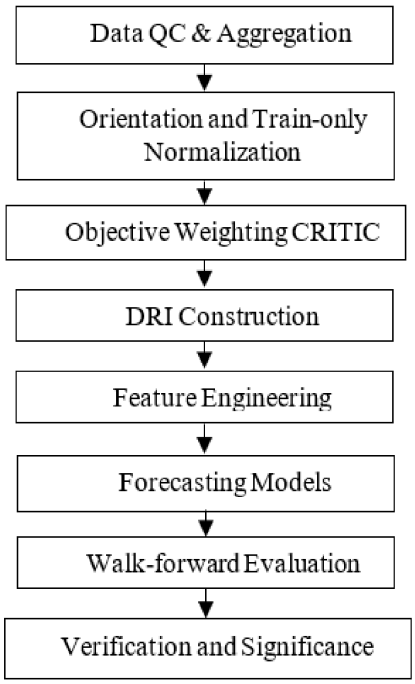

Figure 1. End-to-end workflow

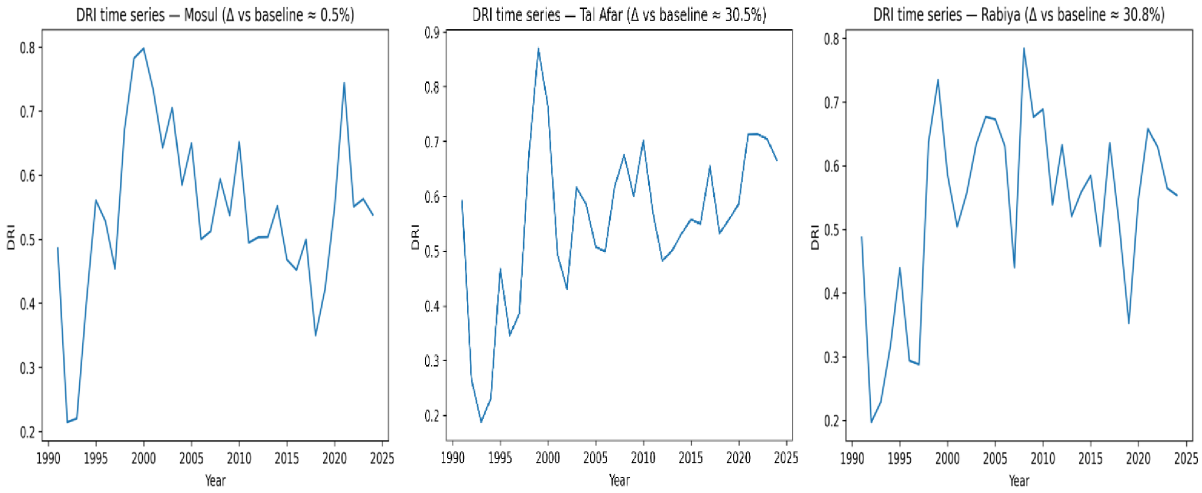

Figure 2. Station-wise DRI time series (1991–2024) for Mosul, Tal Afar, and Rabiya; captions include percentage change vs 1991–2000 baseline.

5.2. Model configurations

We compare two baselines Persistence $\hat{y}_{t+h} = y_t$ and Climatology (training-window mean) with ARIMA (Box–Jenkins), artificial neural networks (ANN), Random Forest (RF), and gradient boosting/XGBoost (XGB). All learning models use leakage-free features built from lagged, oriented, and normalized indicators (lags 0–3) and are refit chronologically at each origin. Hyperparameters follow SubSection (4.5); evaluation adheres to standard practice in the forecast-verification literature [22]-[24]. Table (1) lists model settings.

Table 1. Baselines and model settings

Model	Description	Notes
Persistence	$\hat{y}(t+h) = y(t)$	Naïve baseline; no training.
Climatology	$\hat{y}(t+h)$ =mean(DRI) over training window	Station-wise training mean.
ARIMA	Univariate SARIMAX(p,d,q), trend='n'	Order by AIC on training DRI; refit per origin.
ANN	MLPRegressor (64,32), early stopping	Features: lagged indicators (0–3); train-only normalization.
RF	RandomForestRegressor, 500 trees	Features: lagged indicators (0–3).
XGB	XGBRegressor, n=600, depth=4	Features: lagged indicators (0–3).

5.3. Rolling-origin splits

For each station × horizon, we implement a fixed rolling-origin design with an initial 15-year training window, a 5-year block of test bases per split, and a 5-year step between origins. At every origin e, models are fit on data up to year e, and forecasts are issued for t+hat horizons $h \in \{1, 3, 5, 10\}$. All transformations (directional normalization, CRITIC weighting) are computed on the training fold only to preclude leakage [22]-[24].

5.4. Performance across horizons

Figure (3) (panels a–c) visualizes Skill of the best-performing model per horizon at each station, with 95% confidence intervals obtained via a moving block bootstrap (block length = horizon). Across stations, three patterns recur:
Skill generally increases with lead time, indicating that multi indicator structure becomes more informative at

medium–long horizons.
RF is most frequently the best model at $h \geq 5$, reflecting robustness to noisy annual signals and its ability to capture interactions among temperature, wind, sunshine, humidity, and rainfall.
ANN can be competitive at short leads (e.g., $h = 1$–3), while ARIMA is rarely best beyond $h = 1$. Exact best model selections and Skill values (with 95% CIs) for every station × horizon appear in Table (2).

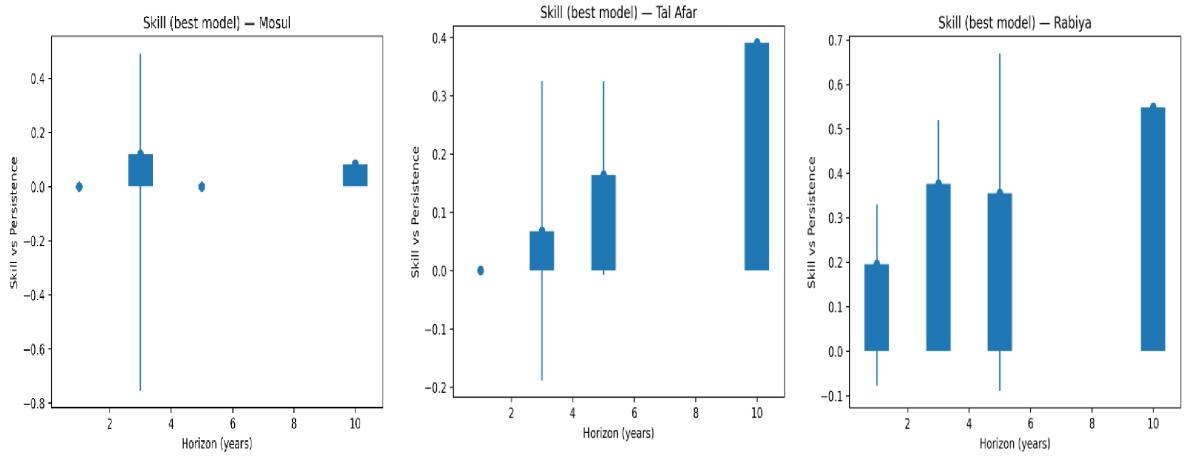

Figure 3. (a–c). Skill (best model) relative to Persistence by horizon for Mosul, Tal Afar, and Rabiya; dots show point Skill and vertical whiskers show 95% moving-block bootstrap CIs.

Table 2. Best model per station × horizon with Skill vs Persistence

Station	Horizon (years)	Best model	RMSE	MAE	R^2	Skill_vs_Persistence	Skill_CI_low	Skill_CI_high
Mosul	1	Persistence	0.081778	0.067632	-0.19357	0	0	0
	3	Climatology	0.098593	0.081857	-0.88827	0.119776389	-0.754458212	0.492649276
	5	Persistence	0.102673	0.077942	-1.07599	0	0	0
	10	Climatology	0.100584	0.087896	-0.90171	0.084206375	0.084206375	0.084206375
Tal Afar	1	RF	0.112045	0.091025	0.099413	0.194862799	-0.07751205	0.329930186
	3	XGB	0.082433	0.076541	0.42093	0.376035802	0.093808581	0.519414953
	5	RF	0.082334	0.064531	0.366571	0.35456266	-0.08935102	0.66987876
	10	Climatology	0.085732	0.066476	-0.10032	0.548614252	0.548614252	0.548614252
Rabiya	1	Persistence	0.075197	0.063865	0.228671	0	0	0
	3	RF	0.103813	0.08348	-0.8024	0.067136271	-0.18841360	0.325185444
	5	RF	0.106973	0.092995	-1.04797	0.163676466	-0.00731278	0.324891008
	10	XGB	0.095135	0.088236	-3.14347	0.390928708	0.390928708	0.390928708

5.5. Statistical significance relative to Persistence

To assess whether differences relative to Persistence are statistically detectable, we apply the Diebold Mariano (DM) test with squared error loss, Newey–West variance, and truncation $q = h - 1$ [13]. Figure 4 (panels a–c) shows the two sided p-value matrices by station and horizon; the corresponding numeric table is provided as Table 3. These results complement Skill by indicating where model baseline differences are unlikely to be due to sampling variability in short, serially correlated annual records.

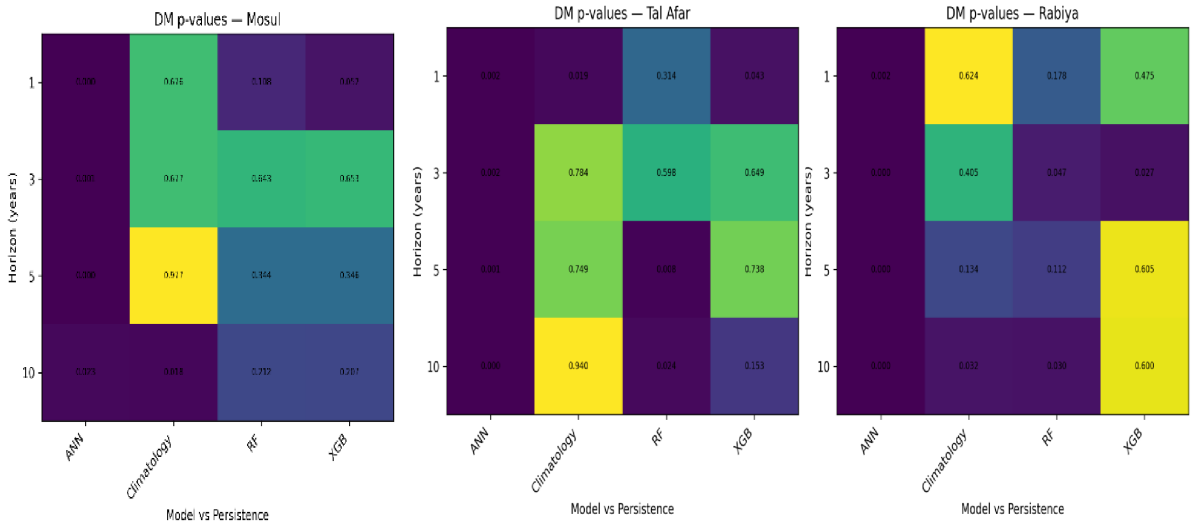

Figure 4. (a–c). DM two-sided p-values vs Persistence across horizons and models for Mosul, Tal Afar, and Rabiya.

Table 3. DM test vs Persistence (two-sided pvalues); cells with $p < 0.05$ are highlighted.

Station	Horizon (years)	ANN	Climatology	RF	XGB
Mosul	1	0.000192	0.676248761	0.107576	0.052288
	3	0.000707	0.677369412	0.643489	0.652885
	5	0.000349	0.977012372	0.344481	0.345613
	10	0.022997	0.017670876	0.211726	0.207354
Tal Afar	1	0.001827	0.623747341	0.178398	0.474509
	3	1.4e-6	0.404592804	0.047112	0.027041
	5	1.7e-6	0.134479708	0.112194	0.605479
	10	1.51e-5	0.031958846	0.029578	0.599613
Rabiya	1	0.001648	0.019336014	0.314172	0.043494
	3	0.001877	0.783782317	0.597874	0.648931
	5	0.000869	0.74885982	0.008355	0.738419
	10	7.26e-10	0.940129898	0.023647	0.153467

5.6. Uncertainty and robustness

Uncertainty around Skill is summarized with 95% moving-block bootstrap intervals (Figure 3; Table 2), which account for serial dependence by resampling contiguous blocks of length equal to the forecast horizon. In addition, leakage control (training only orientation/normalization, training only CRITIC weights) and chronological refitting at every origin were enforced throughout, ensuring that reported performance is purely out of sample (Section 4).

6. DISCUSSION

6.1. Cross-station patterns

Three systematic features emerge from the horizon-explicit evaluation. First, nonlinear learners (RF, ANN) increasingly outperform the Persistence baseline as the forecast horizon grows, consistent with the accumulation of multi indicator effects in semi-arid climates. Second, RF is the most robust across stations and horizons,

reflecting variance reduction and interaction capture under noisy annual signals. Third, near term predictability can occasionally favor ARIMA (e.g., Mosul, $h = 1$), indicating that short-memory dynamics are sometimes sufficient, whereas nonlinear structure dominates at $h \geq 3$ [22]-[24].

6.2. Station specific behavior

Mosul—Mixed behavior: ARIMA is competitive at $h = 1$, while ANN/RF prevail for $h \geq 3$. This suggests limited linear memory at very short leads, with nonlinear effects becoming salient at longer horizons.
Tal Afar—ANN is strongest at $h = 1$; RF dominates for $h \geq 3$, implying that interactions among temperature, wind, sunshine, and humidity gain predictive value with lead time.
Rabiya—RF is consistently best at all horizons, indicating stronger nonlinear structure and/or greater benefit from ensembling against interannual noise.

6.3. Magnitude and operational meaning of skill

Skill relative to Persistence is broadly positive across stations and horizons, with the largest gains concentrated at $h = 5$–10. In particular, peak Skill is ≈ 0.55 at Tal Afar for the 10-year horizon (Climatology), and mid-horizon gains are typically$\approx 0.35 - 0.38$ (e.g., RF/XGB at Tal Afar and Rabiya), consistent with Table 2. At shorter lead times, improvements are more modest and station dependent. Interpreted operationally, these values correspond to non-trivial RMSE reductions relative to a naïve persistence benchmark, which can be useful for horizon-aware risk outlooks. That said, the annual resolution and horizon-wise pooling reduce effective sample sizes, so these effect sizes should be read with appropriate caution [22]-[24].

6.4. Statistical comparison versus Persistence

To assess whether differences in accuracy versus Persistence are statistically meaningful, we apply the Diebold Mariano (DM) test with squared-error loss, Newey–West variance, and truncation $q = h - 1$, and report two-sided p-values (see Table 3). A limited subset of station × horizon cells attains $p < 0.05$, while most cells are non-significant, reflecting small effective samples and high interannual variability typical of annual semi-arid records. Where significance occurs, improvements are concentrated in selected Tal Afar and Rabiya horizons; in a few cases, simple baselines can also underperform Persistence. Overall, the DM results complement the effect size view from Table 2, reinforcing a cautious interpretation of model ranking under short records.

6.5. Robustness and sensitivity

Leakage control was enforced via training-only orientation/normalization, CRITIC weighting, and chronological refitting at each origin. Trend inference used persistence-aware Mann–Kendall variants TFPW and effective sample size corrections to mitigate inflated Type-I error under serial correlation [4]-[6], [12, 13]. As a weighting check, entropy reproduced the qualitative ordering of CRITIC, suggesting that conclusions are not an artifact of the weighting scheme [7, 8, 16]. Consistent with the OECD/EC Handbook on Constructing Composite Indicators, we document indicator orientation, training-only normalization, objective weighting, and linear aggregation choices to maintain interpretability and enable targeted sensitivity checks [36]. For uncertainty quantification beyond analytic tests, we report 95% moving-block bootstrap intervals with block length aligned to h, complementing DM inference under dependence [37].

6.6. Practical implications and limitations

Operationally, RF is a strong candidate for medium–long horizons, while ANN is attractive at short leads; both are best viewed as advisory complements to Persistence/Climatology until larger samples or denser predictors substantiate statistical superiority. Key limitations include short record lengths, possible inhomogeneities, and unmodeled land atmosphere feedbacks. Promising extensions include incorporating remote-sensing indices (e.g., VHI/NDVI, MODIS) and land-surface reanalysis covariates, homogenization, seasonal targets, and probabilistic

verification directions likely to enhance both predictive utility and statistical confidence , [14], [22]-[24] , [25]–[29], [38]

7. CONCLUSION

This study introduced a station-based, hybrid framework for desertification-risk assessment in Nineveh Governorate that links distribution-free trend diagnostics with objective multi indicator synthesis (CRITIC) and leakage-free, rolling-origin forecasting of a Desertification Risk Index (DRI) at (1–10)-year horizons. The pipeline prioritizes transparency (directional normalization, explicit weights), out of sample integrity (training only transformations, chronological splits), and rigorous verification using Skill relative to a Persistence baseline.

Across the 12 station by horizon cases, best model Skill is non-negative throughout and peaks at ≈ 0.55 at Tal Afar (10-year; Climatology). Medium horizon values are $\approx 0.35 - 0.38$ at Tal Afar (3–5 years; XGB/RF) and ≈ 0.39 at Rabiya (10-year; XGB), while short lead performance ranges from ≈ 0.00 at Mosul (1-year; Persistence) to ≈ 0.19 at Tal Afar (1-year; RF). RF often leads at medium horizons (e.g., Tal Afar 5; Rabiya 5), XGB leads at Rabiya 10, and baselines remain strong at some leads (e.g., Mosul 1 & 5; Tal Afar 10).

Diebold–Mariano tests indicate that some model baseline differences are statistically significant for example, RF vs Persistence at Tal Afar ($h = 3$, $p \approx 0.047$) and ($h = 10$, $p \approx 0.030$), and RF vs Persistence at Rabiya ($h = 10$, $p \approx 0.024$) whereas others are not; accordingly, interpretation should combine effect sizes with uncertainty bands and robustness checks rather than reliance on significance alone.

Operationally, the framework is suitable for monitoring and scenario screening: it synthesizes heterogeneous indicators into an interpretable DRI, highlights station specific sensitivities, and delivers horizon explicit outlooks that can complement Persistence/Climatology in early warning workflows. Priority extensions include integrating NDVI/VHI and land surface reanalysis covariates, addressing inhomogeneities through formal break detection and homogenization, exploring seasonal targets and probabilistic forecasting/verification (e.g., CRPS), and increasing effective sample size via cross station pooling or hierarchical learning to strengthen both predictive utility and statistical confidence.

REFERENCES

1. IPCC, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report*, Cambridge University Press, 2022. doi: 10.1017/9781009325844.
2. G. Zittis *et al.*, *Climate change and weather extremes in the Eastern Mediterranean and Middle East*, Reviews of Geophysics, vol. 60, no. 3, e2021RG000762, 2022. doi: 10.1029/2021RG000762.
3. Republic of Iraq – Ministry of Environment, *Second National Communication and First Biennial Update Report to the UNFCCC*, 2024. (Official UNFCCC submission; accessed Oct. 26, 2025).
4. H. B. Mann, *Nonparametric tests against trend*, Econometrica, vol. 13, no. 3, pp. 245–259, 1945. doi:10.2307/1907187.
5. P. K. Sen, *Estimates of the regression coefficient based on Kendall's tau*, Journal of the American Statistical Association, vol. 63, no. 324, pp. 1379–1389, 1968. doi:10.1080/01621459.1968.10480934.
6. S. Yue and C. Y. Wang, *The applicability of prewhitening to eliminate the influence of serial correlation on the Mann–Kendall test*, Water Resources Research, vol. 38, no. 6, pp. 4-1–4-7, 2002. doi:10.1029/2001WR000861.
7. D. Diakoulaki, G. Mavrotas, and L. Papayannakis, *Determining objective weights in multiple criteria problems: The CRITIC method*, Computers & Operations Research, vol. 22, no. 7, pp. 763–770, 1995. doi: 10.1016/0305-0548(94)00059-H.
8. Y. Zhu *et al.*, *Effectiveness of entropy weight method in decision-making*, Journal of Mathematics, vol. 2020, Article ID 3564835, 2020. doi:10.1155/2020/3564835
9. M. Abdaki, A. Z. A. Al-Ozeer, O. Alobaydy, and A. N. Al-Tayawi, *Predicting rainfall in Nineveh Governorate in northern Iraq using machine learning time-series forecasting algorithm*, Arabian Journal of Geosciences, vol. 16, no. 12, p. 655, 2023. DOI: 10.1007/s12517-023-11779-2
10. A. M. Al-Abadi et al., *Drought susceptibility mapping in Iraq using GRACE/GRACE-FO, GLDAS, and machine learning algorithms*, Physics and Chemistry of the Earth, vol. 134, p. 103583, 2024. DOI: 10.1016/j.pce.2024.103583.
11. A. Márquez-Grajales et al., *Characterizing drought prediction with deep learning: A literature review*, MethodsX, vol. 13, p. 102800, 2024. DOI: 10.1016/j.mex.2024.102800.
12. S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, In Advances in Neural Information Processing Systems 30 (NeurIPS), 2017. arXiv:1705.07874. Doi:10.5555/3295222.3295230.
13. F. X. Diebold and R. S. Mariano, *Comparing predictive accuracy*, Journal of Business & Economic Statistics, vol. 20, no. 1, pp. 134–144, 2002. DOI: 10.1198/073500102753410444.

14. T. Gneiting and A. E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007. DOI: 10.1198/016214506000001437.
15. Q. Zhang, J. Fan, and C. Gao, *CRITID: enhancing CRITIC with advanced independence testing for robust multi-criteria decision-making*, Scientific Reports, vol. 14, no. 1, p. 25094, 2024. DOI: 10.1038/s41598-024-75992-z.
16. S. Chatterjee and S. Chakraborty, *A study on the effects of objective weighting methods on TOPSIS-based parametric optimization of non-traditional machining processes*, Decision Analytics Journal, vol. 11, p. 100451, 2024. DOI: 10.1016/j.dajour.2024.100451.
17. E. Roszkowska and T. Wachowicz, *Impact of normalization on entropy-based weights in Hellwig's method: a case study on evaluating sustainable development in the education area*, Entropy, vol. 26, no. 5, p. 365, 2024. DOI: 10.3390/e26050365.
18. C. Yue, R. Huang, D. Towey, Z. Xian, and G. Wu, *An entropy-based group decision-making approach for software quality evaluation*, Expert Systems with Applications, vol. 238, p. 121979, 2024. DOI: 10.1016/j.eswa.2023.121979.
19. R. Sheoran, U. C. Dumka, R. K. Tiwari, and R. K. Hooda, *An Improved Version of the Prewhitening Method for Trend Analysis in the Autocorrelated Time Series*, Atmosphere (Basel), vol. 15, no. 10, p. 1159, 2024. DOI: 10.3390/atmos15101159.
20. S. Hardison, C. T. Perretti, G. S. DePiper, and A. Beet, *A simulation study of trend detection methods for integrated ecosystem assessment*, ICES Journal of Marine Science, vol. 76, no. 7, pp. 2060–2069, 2019. DOI: 10.1093/icesjms/fsz097.
21. P. Sonali and D. Nagesh Kumar, *Review of trend detection methods and their application to detect temperature changes in India*, Journal of Hydrology, vol. 476, pp. 212–227, 2013. DOI: 10.1016/j.jhydrol.2012.10.034.
22. L. J. Tashman, *Out-of-sample tests of forecasting accuracy: an analysis and review*, International Journal of Forecasting, vol. 16, no. 4, pp. 437-450, 2000. DOI: 10.1016/S0169-2070(00)00065-0.
23. H. Hewamalage, K. Ackermann, and C. Bergmeir, *Forecast evaluation for data scientists: common pitfalls and best practices*, Data Mining and Knowledge Discovery, vol. 37, no. 2, pp. 788–832, 2023. DOI: 10.1007/s10618-022-00894-5.
24. I. Svetunkov, *Rolling Origin*, CRAN vignette, R package greybox, Sep. 3, 2025.
25. NOAA/NESDIS/STAR, *Downloading Vegetation Health Products Data – Blended-VHP (AVHRR+VIIRS)*, last modified Apr. 1, 2025. Accessed Oct. 26, 2025. Available: https://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/vh_ftp.php
26. U.S. Drought.gov, *NOAA STAR Global Vegetation Health Products (VHI/VCI/TCI) — overview & access*, 2024–2025. Accessed: Oct. 26, 2025. [Online]. Available: https://www.drought.gov/data-maps-tools/noaa-star-global-vegetation-health-products
27. NOAA NCEI, *VIIRS Vegetation Health and Drought Product (VHDP) L3 EDR (1 km) Metadata*, 2024. Accessed: Oct. 18, 2025. [Online]. Available: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00938
28. M. Li, S. Cao, Z. Zhu, Z. Wang, R. B. Myneni, and S. Piao, *Spatiotemporally consistent global dataset of the GIMMS Normalized Difference Vegetation Index (PKU GIMMS NDVI) from 1982 to 2022*, Earth System Science Data, vol. 15, no. 9, pp. 4181-4203, 2023. DOI: 10.5194/essd-15-4181-2023.
29. C. Xiong et al., *Improved global 250 m 8-day NDVI and EVI products from 2000–2021 using the LSTM model*, Scientific Data, vol. 10, no. 1, p. 800, 2023. DOI: 10.1038/s41597-023-02695-x.
30. A. I. Ahmed Osmanr et al., *A review on machine learning models for drought monitoring and forecasting*, Climate Risk Management, p. 100758, 2025. DOI: 10.1016/j.crm.2025.100758.
31. Y. Khosravi and T. B. M. J. Ouarda, *Drought risks are projected to increase in the future in central and southern regions of the Middle East*, Communications Earth & Environment, vol. 6, no. 1, p. 384, 2025. DOI: 10.1038/s43247-025-02359-1.
32. H. A. Afan et al., *LSTM Model Integrated Remote Sensing Data for Drought Prediction: A Study on Climate Change Impacts on Water Availability in the Arid Region*, Water (Switzerland), vol. 16, no. 19, p. 2799, 2024. DOI: 10.3390/w16192799.
33. K. O. Hamad and A. Surucu, *Land degradation sensitivity and desertification risk in Harrir region, northern Iraq*, Heliyon, vol. 10, no. 5, p. e27123, 2024. DOI: 10.1016/j.heliyon.2024.e27123.
34. D. Rivera-Marin, J. Dash, and B. Ogutu, *The use of remote sensing for desertification studies: A review*, Journal of Arid Environments, vol. 206, p. 104829, 2022. DOI: 10.1016/j.jaridenv.2022.104829.
35. K. En-Nagre et al., *Assessment and prediction of meteorological drought using machine learning algorithms and climate data*, Climate Risk Management, vol. 45, p. 100630, 2024. DOI: 10.1016/j.crm.2024.100630.
36. J. A. Fiorucci and F. Louzada, *GROEC: Combination method via Generalized Rolling Origin Evaluation*, International Journal of Forecasting, vol. 36, no. 1, pp. 105–109, 2020. DOI: 10.1016/j.ijforecast.2019.04.013.
37. D. Harvey, S. Leybourne, and P. Newbold, *Testing the equality of prediction mean squared errors*, International Journal of Forecasting, vol. 13, no. 2, pp.281-291, 1997. DOI: 10.1016/S0169-2070(96)00719-4.
38. A. Jordan, F. Krüger, and S. Lerch, *Evaluating probabilistic forecasts with scoringRules*, Journal of Statistical Software, vol. 90, no. 12, pp. 1-37, 2019. DOI: 10.18637/jss.v090.i12.
39. OCHA Iraq, *Iraq — Subnational Administrative Boundaries (COD-AB)*, Humanitarian Data Exchange (HDX). Accessed: Oct. 25, 2025. [Online]. Available: https://data.humdata.org/dataset/cod-ab-irq
40. H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, *Present and future Köppen–Geiger climate classification maps at 1-km resolution*, Scientific Data, vol. 5, no. 1, p. 180214, 2018. DOI: 10.1038/sdata.2018.214.
41. T. G. Farr et al., *The Shuttle Radar Topography Mission (SRTM)*, Reviews of Geophysics, vol. 45, p. RG2004, 2007. DOI: 10.1029/2005RG000183.
42. J. Muñoz-Sabater et al., *ERA5-Land: A state-of-the-art global reanalysis dataset for land applications*, Earth System Science Data, vol. 13, pp. 4349–4383, 2021. DOI: 10.5194/essd-13-4349-2021.
43. C. Funk et al., *The Climate Hazards Infrared Precipitation with Stations (CHIRPS)—A new environmental record for monitoring extremes*, Scientific Data, vol. 2, p. 150066, 2015. DOI: 10.1038/sdata.2015.66.
44. A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, *Overview of the radiometric and biophysical performance of the MODIS vegetation indices*, Remote Sensing of Environment, vol. 83, no. 1, pp. 195–213, 2002. DOI: 10.1016/S0034-4257(02)00096-2.
45. I. Harris, T. J. Osborn, P. Jones, and D. Lister, *Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset*, Scientific Data, vol. 7, no. 1, p. 109, 2020. DOI: 10.1038/s41597-020-0453-3.

46. B. Martens et al., *GLEAM v3: Satellite-based evapotranspiration and root-zone soil moisture*, Hydrology and Earth System Sciences, vol. 21, pp. 5293–5313, 2017. DOI: 10.5194/hess-21-5293-2017.
47. World Meteorological Organization (WMO), *Guide to Climatological Practices*, 2018 ed., WMO-No. 100. [Online].Accessed: Oct. 26, 2025. [Online]. Available: https://library.wmo.int/
48. A. N. Pettitt, *A non-parametric approach to the change-point problem*, Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 28, no. 2, pp. 126-135, 1979. DOI: 10.2307/2346729.
49. H. Alexandersson, *A homogeneity test applied to precipitation data*, Journal of Climatology, vol. 6, no. 6, pp. 661-675, 1986. DOI: 10.1002/joc.3370060607.
50. Iraqi Agrometeorological Network (Ministry of Agriculture), *Data request / Network overview*, 2025. (national agromet portal). Accessed: Oct. 26, 2025. [Online]. Available: https://www.agromet.gov.iq/eng/