

Quantifying Seasonal Vegetation Anomalies in an Arid Ecosystem: A Reproducible, Leakage-Free Machine Learning Framework with Dynamic Climate Predictors and Quantified Uncertainty

aamna naveed

aamnat340@gmail.com

Research Article

Keywords: arid ecosystems, vegetation stress, MODIS NDVI, machine learning, spatial cross-validation, data leakage, uncertainty quantification, Cholistan Desert, NASA POWER

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11188902/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantifying Seasonal Vegetation Anomalies in an Arid Ecosystem: A Reproducible, Leakage-Free Machine Learning Framework with Dynamic Climate Predictors and Quantified Uncertainty

Aamna Naveed

Independent Researcher, Bahawalpur, Pakistan

Correspondence: aamnat340@gmail.com

ORCID: 0009-0001-3371-1066

Abstract

Vegetation stress in arid ecosystems is a critical early-warning indicator for desertification and agricultural risk, yet its modelling remains methodologically challenging: climate forcing is highly stochastic, and standard anomaly pipelines frequently leak future information into historical baselines, inflating apparent predictive skill. This study presents a reproducible, open-source machine-learning framework for modelling seasonal vegetation anomalies in the Cholistan Desert, Pakistan, using MODIS NDVI (MOD13Q1; 250 m, 16-day; 2018–2024). In contrast to conventional workflows built on static climate normals, we derive time-varying seasonal predictors (precipitation, one-season-lagged precipitation as a proxy for antecedent moisture conditions, and temperature) from the NASA POWER archive, alongside SRTM terrain derivatives and MODIS land cover (MCD12Q1). To guarantee an honest evaluation, the seasonal anomaly baseline (pixel- and season-specific mean and standard deviation) is estimated strictly on the 2018–2023 training period, and the entire year 2024 is withheld. Generalized additive models (GAM) and random forests (RF) are compared under 4-fold spatial-block cross-validation ($n = 30,646$ valid pixel-composite observations assigned to meteorological seasons) and the temporal hold-out ($n = 4,179$ valid pixel-composite observations from the 2024 hold-out). RF consistently outperformed GAM (mean spatial-CV = 0.261 vs. 0.180; 2024 hold-out = 0.156 vs. 0.100), with seasonal temperature, seasonal precipitation, and lagged precipitation dominating variable importance, supporting the use of dynamic climate predictors. Predictive uncertainty, proxied by GAM–RF disagreement, was markedly elevated where 2024 climate conditions fell outside the training distribution, transparently exposing where predictions amount to extrapolation. The framework provides a leakage-free template for dryland vegetation monitoring under intensifying climate variability and identifies priority spatiotemporal domains for targeted field validation.

Keywords: arid ecosystems; vegetation stress; MODIS NDVI; machine learning; spatial cross-validation; data leakage; uncertainty quantification; Cholistan Desert; NASA POWER

1. Introduction

Dryland ecosystems cover more than 40% of the Earth's land surface and support over two billion people, yet they are among the environments most sensitive to climatic variability, extreme heatwaves, and erratic precipitation [1]. Monitoring vegetation condition in these regions is essential for early warning of desertification, for rangeland and agricultural risk assessment, and for guiding ecological restoration. Satellite remote sensing, above all the Normalized Difference Vegetation Index (NDVI) from the Moderate Resolution Imaging Spectroradiometer (MODIS), provides a continuous, freely available, multi-decadal record of vegetation greenness at regional scales [2,3].

Despite the abundance of satellite data, modelling vegetation stress in arid regions faces three persistent methodological shortcomings. *First*, many studies use static, long-term climate averages (e.g., WorldClim bioclimatic variables) as predictors [4]. In hyper-arid systems, however, vegetation responds to episodic rainfall events and heatwaves rather than to long-run climatic normals; static averages discard precisely the temporal signal that drives vegetation response [5]. *Second*, conventional anomaly calculations are vulnerable to **temporal data leakage**: when the full time series, including the evaluation period, is used to estimate the historical mean and standard deviation, test information contaminates the baseline and model performance is artificially inflated [6]. *Third*, flexible machine-learning models are routinely deployed without rigorous uncertainty quantification or out-of-distribution diagnostics, leaving end users unable to distinguish trustworthy predictions from silent extrapolation beyond the training domain [7].

These shortcomings are especially consequential in data-scarce drylands such as the Cholistan Desert of southern Punjab, Pakistan, where sparse ground-observation networks make satellite-based monitoring the only viable continuous information source, and where methodological shortcuts can quietly compound into misleading management advice.

This study addresses these gaps through a strictly leakage-free, fully reproducible R-based workflow, with three specific contributions:

1. a seasonal NDVI anomaly response whose baseline is estimated **exclusively on the training period**, eliminating temporal leakage by construction;
2. the integration of **dynamic, time-varying climate predictors** (seasonal precipitation, one-season-lagged precipitation, and seasonal temperature) together with genuine MODIS land cover, replacing static climate normals; and
3. a validation and uncertainty framework combining **spatial-block cross-validation, a strict temporal hold-out, model-disagreement diagnostics, and environmental-novelty screening**, so that predictive reliability is quantified rather than assumed.

Accordingly, the objectives of this study are to (i) quantify the seasonal vegetation anomaly dynamics of the Cholistan Desert from 2018 to 2024; (ii) compare the predictive skill of GAM and RF under spatially and temporally honest validation; and (iii) map where, and under what climatic conditions, model predictions cease to be reliable.

2. Materials and Methods

2.1 Study area

The study area comprises the Cholistan Desert region of southern Punjab, Pakistan, encompassing the districts of Bahawalpur, Bahawalnagar, and Rahim Yar Khan (Figure 1). The region has an extreme continental desert climate: summer temperatures frequently exceed 45 °C, precipitation is low and erratic, and the landscape is dominated by longitudinal sand dunes and interdune plains at elevations of roughly 73–175 m above sea level. Vegetation is sparse and strongly coupled to episodic monsoon rainfall, making the region a demanding test bed for vegetation-stress modelling.

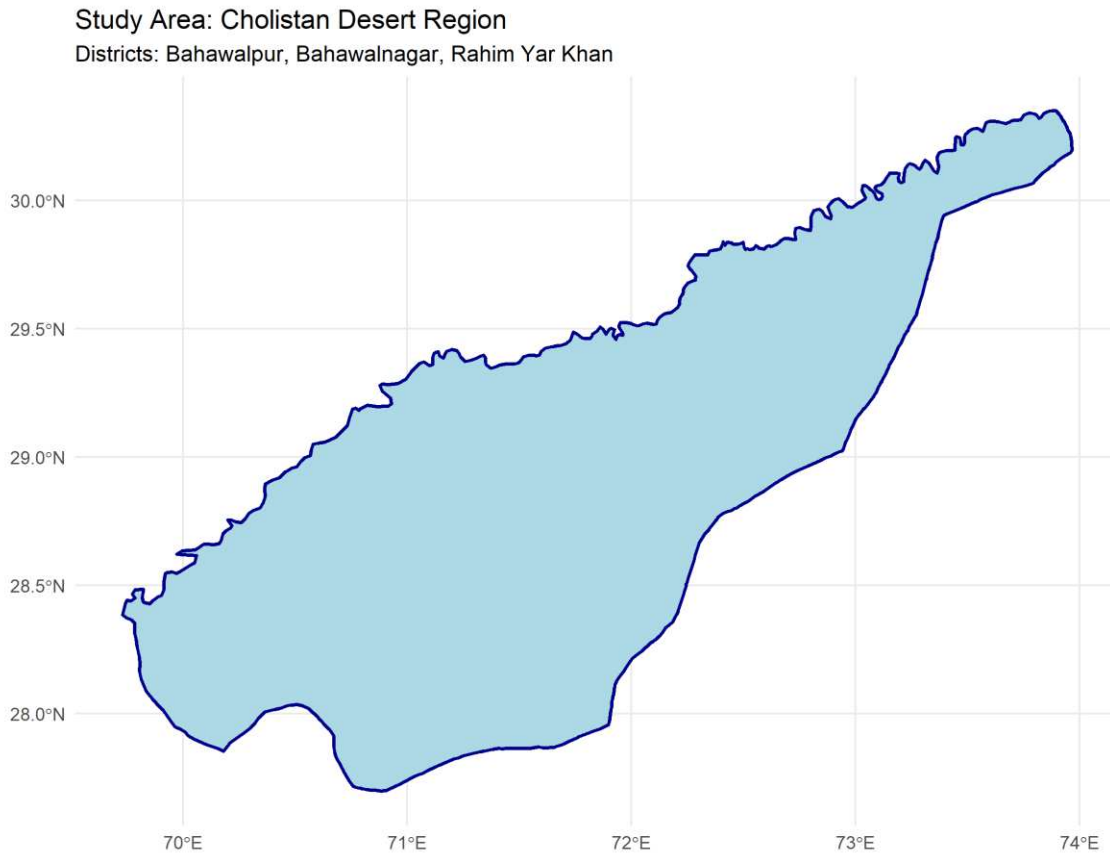


Figure 1. Study area. (a) Location of the Cholistan Desert within Pakistan; (b) the three study districts (Bahawalpur, Bahawalnagar, and Rahim Yar Khan) with SRTM elevation shading and the 200 sampled MODIS pixels overlaid.

2.2 Data acquisition and sampling design

The analysis is built on 200 unique sampling pixels distributed across the three districts generated on a regular grid across the study-area polygon (`sf::st_sample()`, `type = "regular"`; `seed = 42`). Each pixel provides a continuous MODIS NDVI record and collocated climate, terrain, and land-cover predictors, yielding 34,825 valid pixel-composite observations assigned to meteorological seasons in total: 30,646 for 2018–2023 (training and spatial cross-validation) and 4,179 for the withheld year 2024.

- **Vegetation.** MODIS MOD13Q1 Version 6.1 NDVI (250 m, 16-day composite [2,3]) was retrieved for 2018–2024 using the MODISTools R package for each sampled coordinate. Fill values (-3000) were masked as missing. Each pixel and meteorological season could contain several valid 16-day composites; rather than aggregating these into a single seasonal value, every valid composite was retained as an individual observation and assigned to its meteorological season (Section 2.3). This design choice, and its implications for sample size and observation independence, is addressed in Section 4.4.
- **Dynamic climate predictors.** Monthly precipitation and near-surface air temperature were extracted via the NASA POWER API [8] for each sampled coordinate. Precipitation (mm day^{-1}) was aggregated to seasonal totals and temperature (K) converted to $^{\circ}\text{C}$ and averaged per season. A one-season-lagged precipitation predictor (Section 2.3) was additionally computed to proxy antecedent moisture conditions.
- **Terrain.** Elevation was obtained from the SRTM Digital Elevation Model [9]; slope (degrees) was derived with `terra::terrain()`.

- **Land cover.** MODIS MCD12Q1 Version 6.1 (IGBP; 2021) [10] was used as a static categorical predictor. The 17 IGBP classes were recoded into four ecologically meaningful categories for the region: Bare/Sparse, Grass/Shrub, Cropland, and Other.

The final modelling dataset thus contains six predictors per valid pixel–composite observation assigned to a season: seasonal precipitation, one-season-lagged precipitation, seasonal temperature, elevation, slope, and land-cover class.

2.3 Response variable: a leakage-free seasonal anomaly

Season definition. Observations were grouped by meteorological season: Winter (December–February), Spring (March–May), Summer (June–August), and Autumn (September–November). To preserve chronological ordering, December was assigned to the *following* year’s winter (e.g., December 2018 belongs to Winter 2019).

Lagged precipitation. The lagged predictor for a given pixel and season is the total accumulated precipitation of the immediately preceding chronological season. For example, for a pixel in Summer 2020 (JJA), the lagged variable is Spring 2020 (MAM) precipitation; for Winter 2020 (DJF), it is Autumn 2019 (SON) precipitation. This variable captures soil-moisture memory: moisture stored in the soil profile from the previous season, which sustains vegetation beyond the rainfall event itself.

Anomaly construction. To isolate vegetation stress from recurring seasonal phenology, we model the standardized seasonal NDVI anomaly (Z-score). The defining methodological choice is that the seasonal baseline is estimated **strictly on the 2018–2023 training period**, grouped by pixel and season. For pixel i and season s :

$$\mu_{i,s} = \frac{1}{N} \sum_{t=1}^N NDVI_{i,s,t}, \quad t \in [2018, 2023]$$

$$\sigma_{i,s} = \sqrt{\frac{1}{N-1} \sum_{t=1}^N (NDVI_{i,s,t} - \mu_{i,s})^2}$$

The anomaly for any year y , including the withheld 2024 observations, is then:

$$Z_{i,s,y} = \frac{NDVI_{i,s,y} - \mu_{i,s}}{\sigma_{i,s}}$$

Pixels with $\sigma_{i,s} = 0$ or fewer than five valid training observations ($N < 5$) were excluded. Because the 2024 data never enter the baseline, no information from the evaluation period can contaminate model fitting; the pipeline is leakage-free by construction rather than by assumption. Positive anomalies indicate greener-than-normal conditions; negative anomalies indicate vegetation stress relative to the pixel’s own seasonal history.

2.4 Modelling and validation framework

Two complementary learners were compared. A Generalized Additive Model [11] was fitted with a Gaussian family, using thin-plate regression splines with the default basis dimension ($= 10$) for each continuous predictor and a factor effect for land cover, with smoothness parameters estimated by restricted maximum likelihood (REML). A random forest regression [12] was fitted with 500 trees, the default $= 2$ (two candidate predictors considered per split, appropriate for the six-predictor design space), and variable importance measured by mean decrease in impurity.*kmtry*

Validation followed two complementary designs. To mitigate spatial autocorrelation, which inflates skill when nearby observations are split across training and test sets, the 200 sampling coordinates were partitioned into four clusters by K-means clustering on latitude–longitude. Models were iteratively trained on three blocks and tested on the fourth (spatial-block cross-validation), producing test folds of 3,850, 8,778, 7,238, and 10,780 observations. To test temporal transferability, models were then trained on all of 2018–2023 ($n = 30,646$) and evaluated on the fully unseen 2024 data ($n = 4,179$; temporal hold-out). Predictive skill was assessed with root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2), computed as the squared Pearson correlation between observed and predicted values, reported per fold and averaged.

2.5 Uncertainty quantification and environmental novelty

Because point predictions alone conceal model risk in drylands, we quantified uncertainty in two ways. First, predictive disagreement was proxied by the absolute difference between GAM and RF predictions, $|\hat{y}_{\text{GAM}} - \hat{y}_{\text{RF}}|$; large disagreement indicates regions where the data do not constrain the response surface. Second, we screened the withheld 2024 observations for **environmental novelty**: an observation was flagged when any continuous climate predictor fell outside the 5th–95th percentile envelope of the 2018–2023 training distribution. Flagged observations lie beyond the models’ domain of applicability, in the spirit of multivariate environmental similarity surfaces (MESS) [7], and their predictions should be interpreted as extrapolations.

3. Results

3.1 Characteristics of the seasonal anomaly response

The seasonal NDVI anomaly series shows the behaviour expected of a standardized dryland response: the large majority of pixel–seasons cluster near zero (conditions close to the pixel’s historical norm), with asymmetric tails capturing anomalous greenness and anomalous stress (Figure 2). This distributional shape, heavy tails over a near-normal core, reflects the episodic nature of desert vegetation response: most seasons are unremarkable, while a minority of seasons depart sharply from baseline following droughts or unusually effective rainfall.

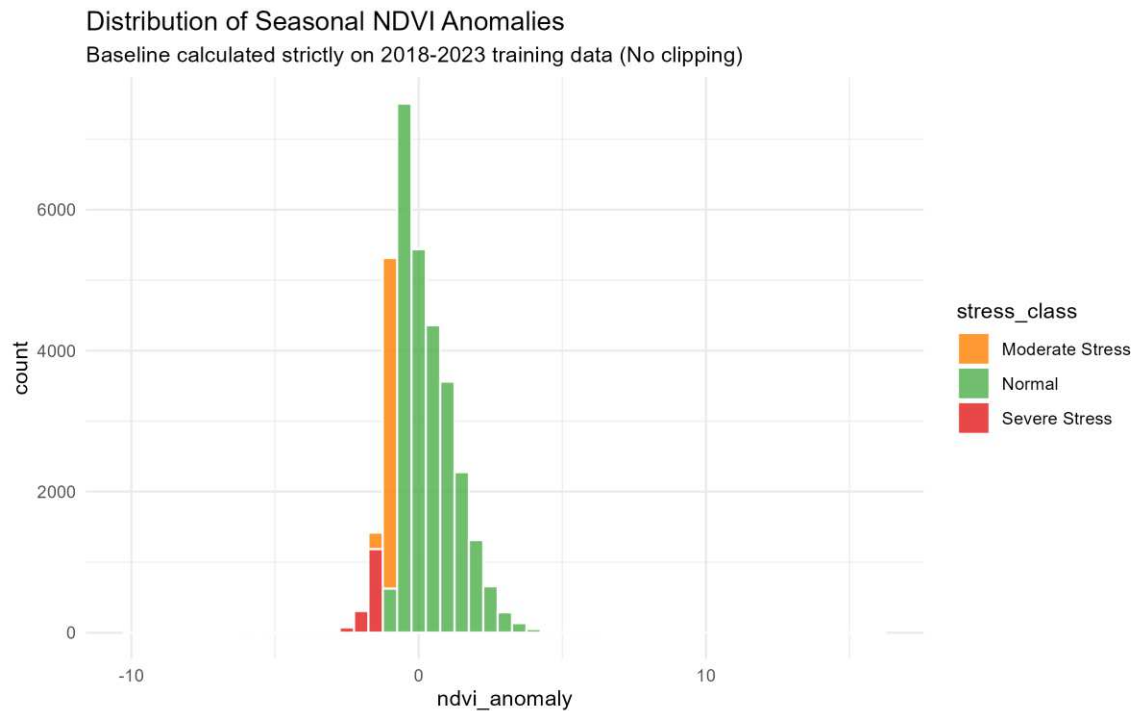


Figure 2. Distribution of seasonal NDVI Z-score anomalies for the 2018–2024 study period.

3.2 Predictive performance under honest validation

Random forest outperformed the GAM under both validation designs (Table 1; Figure 3). In spatial cross-validation, RF achieved a mean R^2 of 0.261 and RMSE of 0.868 across the four spatial blocks, versus 0.180 and 0.921 for the GAM, and this ranking was consistent within each fold, not an artefact of averaging. In the strict 2024 temporal hold-out (the most demanding test, since the models saw neither those observations nor their climatic conditions during training), RF achieved $R^2 = 0.156$ and RMSE = 1.229 on 4,179 unseen observations, versus $R^2 = 0.100$ and RMSE = 1.288 for the GAM. Notably, skill degrades from spatial to temporal validation for both learners (RF R^2 : 0.261 $\rightarrow$ 0.156), exactly the pattern expected when temporal transfer is genuinely harder than spatial interpolation.

While these values are modest in absolute terms, they represent a conservative, leakage-free estimate of transferable skill in a system dominated by unobserved, micro-scale controls such as localized grazing, groundwater depth, and dune morphology. Under this evaluation design, reported skill cannot be an artefact of information leakage, and RF’s consistent advantage is consistent with the presence of non-linear, interacting ecological responses that additive smooths cannot fully represent.

Table 1. Summary of model performance. Spatial-CV metrics are means across the four spatial blocks; fold-level metrics are given in the supplementary material. RMSE and MAE are expressed in anomaly (Z-score) units.

Model	Validation	Test n	Mean RMSE	Mean MAE	Mean R^2
GAM	Spatial CV (4 folds)	3,850–10,780/fold	0.921	0.719	0.180
RF	Spatial CV (4 folds)	3,850–10,780/fold	0.868	0.660	0.261
GAM	Temporal hold-out (2024)	4,179	1.288	0.936	0.100
RF	Temporal hold-out (2024)	4,179	1.229	0.894	0.156

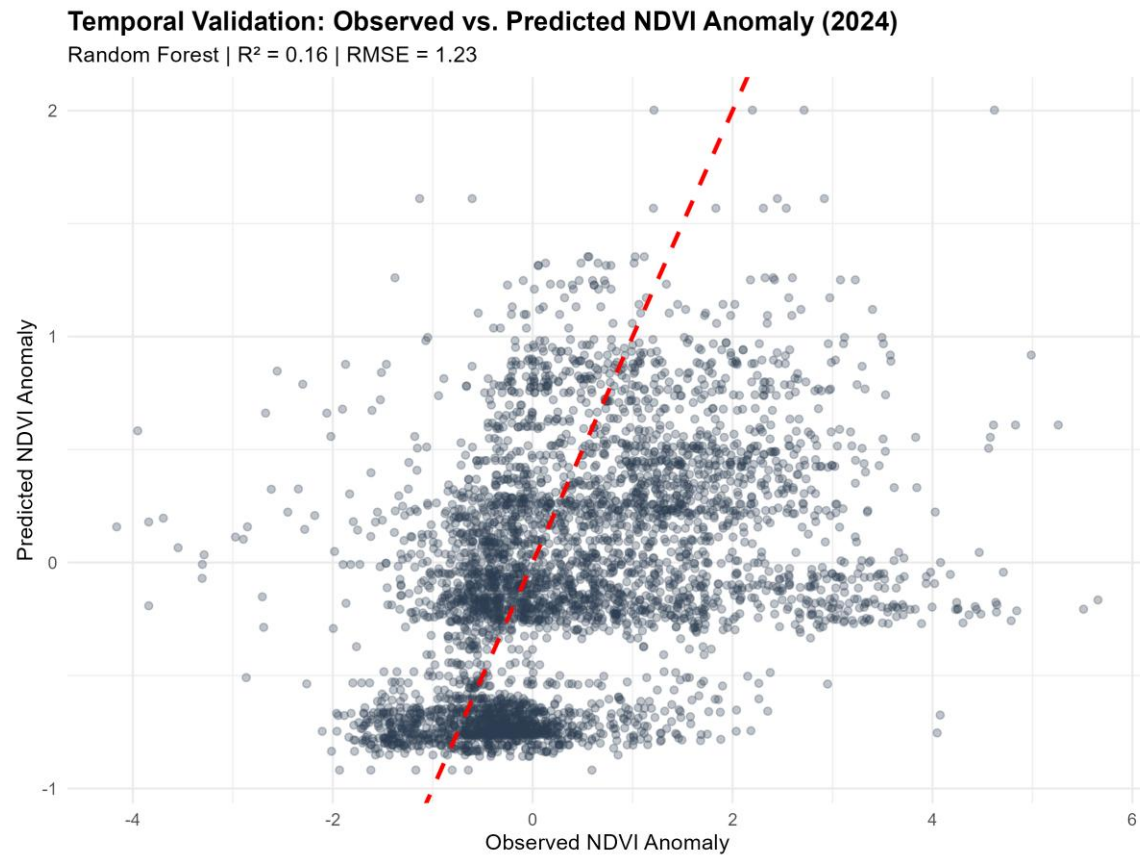


Figure 3. Predicted versus observed seasonal NDVI anomalies on the 2024 temporal hold-out for the Random Forest model, with the 1:1 line shown.

3.3 Variable importance and response shapes

Mean-decrease-impurity importance from the RF identified seasonal temperature as the strongest predictor, followed by seasonal precipitation and one-season-lagged precipitation (Figure 4), supporting the decision to replace static climate normals with time-varying predictors. Terrain variables (elevation, slope) and land-cover class contributed comparatively little predictive information, suggesting that in this flat, low-relief landscape, climate variability overwhelms static structural controls at the 250 m scale.

GAM partial-effect curves reveal the *shape* of these climatic controls (Figure 5). The temperature response is markedly U-shaped: vegetation stress peaks under extreme cold ($< 15^{\circ}\text{C}$) and extreme heat ($> 35^{\circ}\text{C}$), with the most favourable conditions in the intermediate range, consistent with the combined effects of cold-season dormancy and heat stress near the physiological limits of desert flora. Precipitation responses are positive but strongly non-linear, with diminishing returns at high totals, while the lagged-precipitation curve confirms that antecedent moisture carries independent predictive signal beyond same-season rainfall.

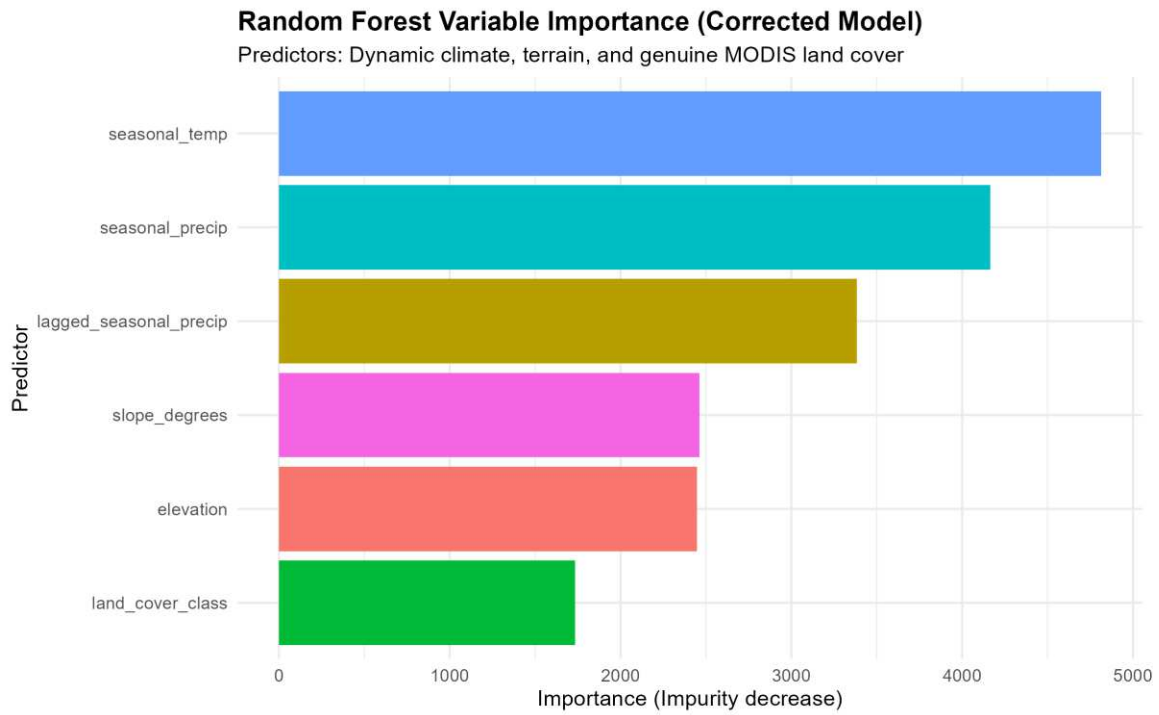


Figure 4. Random-forest variable importance (mean decrease in impurity) across the six predictors.

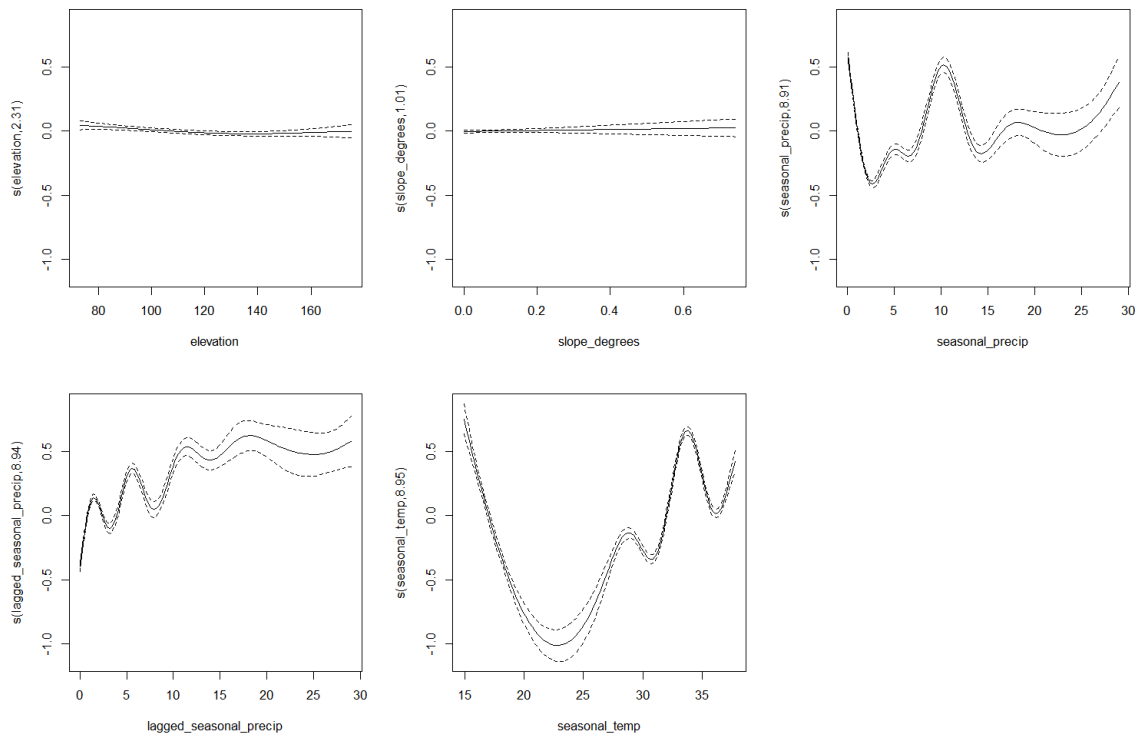


Figure 5. GAM partial-effect curves (thin-plate splines, REML) for the continuous predictors, with 95% confidence limits.

3.4 Uncertainty is concentrated where climate is novel

Model disagreement was higher, on average, among 2024 observations flagged as environmentally novel, indicating increased structural uncertainty under novel climate conditions (Figure 6). This pattern identifies

domains where the models may be extrapolating beyond the central training distribution, but it is an uncertainty proxy rather than a direct estimate of prediction error.

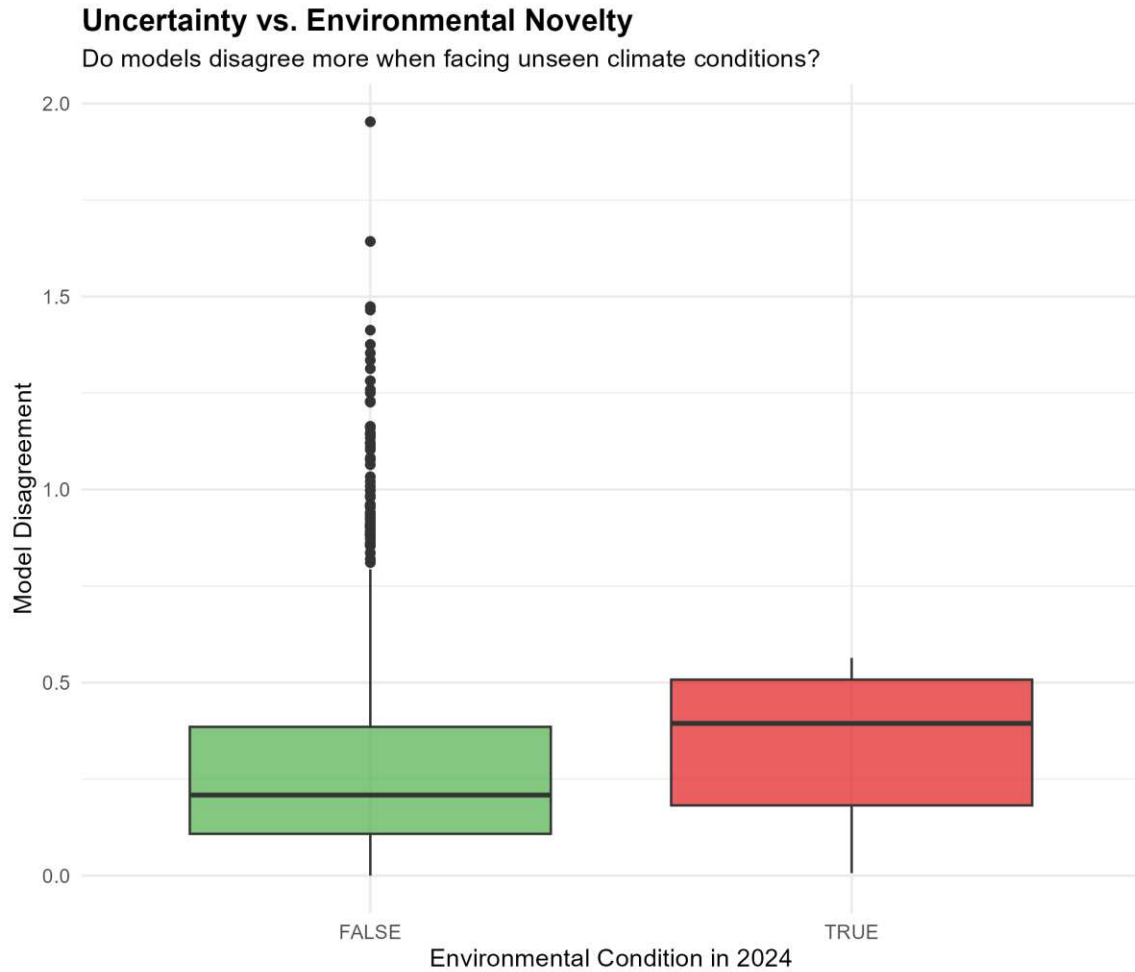


Figure 6. Distribution of model disagreement ($|GAM - RF|$) by environmental-novelty status for 2024 hold-out observations.

4. Discussion

4.1 Antecedent precipitation is strongly associated with dryland vegetation anomalies

The high importance of lagged precipitation is the central ecological finding of this study. Vegetation in the Cholistan Desert does not respond instantaneously to rainfall; it depends on antecedent moisture stored in the soil profile from preceding seasons: moisture that buffers plants against the long dry intervals between the erratic rainfall pulses that characterize the region [5]. Three lines of evidence converge here: lagged precipitation ranks among the top three predictors in the RF; its GAM partial effect is positive and non-trivial in magnitude; and the same lag improves out-of-sample skill under both validation designs. This pattern is consistent with a soil-moisture-memory hypothesis, suggesting that seasonal forecasting and early-warning systems for dryland vegetation must incorporate lagged hydrological predictors, and that static climatologies are structurally incapable of representing the dominant process.

4.2 The honest cost of leakage prevention

The modest R^2 values reported here should be read as a feature of the experimental design, not a failure of the models. By constructing the anomaly baseline exclusively from training-period data and withholding an entire year, we eliminated the most common source of inflated performance in NDVI anomaly studies [6]. Workflows that compute baselines over the full record routinely report substantially higher skill; such numbers reflect memorization of the evaluation period rather than transferable predictive power. The consistent spatial-to-temporal degradation we observe (RF R^2 : 0.261 $\rightarrow$ 0.156) is precisely the signature of honest evaluation: models are always better at interpolating space than at extrapolating time. We argue that the dryland-modelling community should treat leakage-free evaluation as the default standard, and report apparent skill under full-record baselines only as a diagnostic upper bound.

4.3 Practical implications for monitoring and management

The disagreement-versus-novelty pattern provides a practical screening tool: areas where RF and GAM disagree under novel climate conditions are candidate priorities for field validation and ground-truthing, not confirmed stress hotspots. For a region like Cholistan, where ground observations are sparse and budgets for field campaigns are limited, such a triage map directs limited resources to where model-based inference is weakest, and, symmetrically, allows managers to act with greater confidence where disagreement is low. Given projections of intensified heat and rainfall variability for South Asian drylands, the ability to say where and when a model should not be trusted is arguably as valuable as the predictions themselves.

4.4 Limitations and future directions

A methodological limitation concerns the structure of the modelled observations. Rather than aggregating the 16-day MODIS composites within each pixel-season unit into a single seasonal value, every valid composite was retained as a separate observation and paired with the same season-level climate and terrain predictors (Section 2.2). Consequently, the reported sample sizes ($n = 30,646$ training; $n = 4,179$ hold-out) are composite-level rather than season-level, predictor values repeat within a season, and observations are not fully independent; the spatial-block and temporal hold-out metrics in Table 1 should therefore be read as an internally consistent comparison between GAM and RF under this design, not as estimates calibrated to independent seasonal replicates. A stricter one-value-per-pixel-season aggregation, at roughly one-sixth the sample size, is a natural robustness check for future work.

Beyond this, the study quantifies predictive associations, not causal effects; unobserved controls, including groundwater extraction, grazing pressure, and soil texture, remain in the residuals, and in a flat sand-dune landscape micro-topography below the 250 m pixel scale may matter. The environmental-novelty screen is a simplified per-variable quantile approximation rather than a full multivariate MESS implementation, and the 5th–95th envelope is a heuristic rather than a calibrated extrapolation boundary. The 16-day compositing and seasonal aggregation can smooth sub-seasonal stress events, so a pulse visible in weekly data may be averaged away in seasonal means, and in very sparse canopies NDVI remains sensitive to soil-background brightness, which may contribute residual noise. Land cover was held static (2021), although cover change over 2018–2024 may modulate some responses. Finally, no in situ validation was available; flagged high-disagreement/high-novelty domains should be treated as hypotheses for field confirmation. Future work should assimilate Sentinel-2 (10–20 m) imagery to resolve dune-scale heterogeneity, soil-property layers such as SoilGrids to represent texture and water-holding capacity, edaphic and groundwater covariates, and in situ vegetation measurements; formal conformal-prediction intervals should replace the disagreement proxy used here; and the framework should be stress-tested in other drylands to establish the generality of the antecedent-precipitation signal.

5. Conclusion

We developed and validated a reproducible, leakage-free machine-learning pipeline for seasonal vegetation-stress monitoring in the Cholistan Desert. Three design choices define its value: training-period-only anomaly baselines, dynamic time-varying climate forcing, and dual spatial-temporal validation with explicit uncertainty and novelty diagnostics. Across 34,825 valid pixel-composite observations assigned to meteorological seasons, random forest provided the most transferable predictions, temperature and precipitation (including its one-season lag) dominated the signal, and uncertainty concentrated exactly where 2024 climate conditions departed from the historical record, demonstrating that the framework not only predicts but also knows where it should not be trusted. The pipeline is fully open-source and directly transferable to other data-scarce drylands facing intensifying climate variability.

Data and code availability

All data used are freely available: MODIS products via the NASA LP DAAC (MOD13Q1 v6.1; MCD12Q1 v6.1), NASA POWER via its public API, and SRTM via the USGS. The complete analysis pipeline, configuration, and outputs are available at <https://github.com/AamnaNaveed/cholistan-vegetation-stress> (MIT License).

Author contributions

A.N. conceptualized the study, acquired and processed the data, developed the analysis pipeline, performed the analysis, and wrote the manuscript.

Funding

This research received no external funding.

Ethics declarations

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Clinical trial registration: not applicable.

References

- [1] Reynolds, J. F., et al. (2007). Global desertification: building a science for dryland development. *Science*, 316(5826), 847–851. [2] Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2), 195–213. [3] Didan, K. (2021). MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V061. NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD13Q1.061> [4] 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. [5] Ponce Campos, G. E., Moran, M. S., Huete, A., Zhang, Y., Breshears, D. D., Eamus, D., ... & Law, D. J. (2013). Ecosystem resilience despite large-scale altered hydroclimatic conditions. *Nature*, 494(7437), 349–352. [6] Roberts, D. R., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. [7] Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620–1633. [8] NASA Langley Research Center (LaRC) POWER Project (2024). NASA

Prediction of Worldwide Energy Resources (POWER). Accessed via API. <https://power.larc.nasa.gov/> [9] Farr, T. G., et al. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics*, 45(2), RG2004. [10] Sulla-Menashe, D., & Friedl, M. A. (2018). User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C2) Product. NASA. [11] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73(1), 3–36. [12] Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [spatialcvmetrics.csv](#)
- [modelcomparisonsummary.csv](#)
- [temporalholdout2024metrics.csv](#)