Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early Warning of Meteorological Drought in Morocco's Atlas Mountains: A Satellite- Augmented Spatiotemporal Deep Learning Approach

Nacer Aderdour

nacer.aderdour-etu@etu.univh2c

University of Hassan II Casablanca

Hassan Rhinane

University of Hassan II Casablanca

Henri Rueff

University of Bern

Mehdi Maanan

University of Hassan II Casablanca

Research Article

Keywords: drought forecasting, SPI-6, ConvGRU, deep learning, remote sensing, NDVI, Morocco, Atlas Mountains

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9498065/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Theoretical and Applied Climatology on October 10th, 2026. See the published version at https://doi.org/10.1007/s00704-026-06633-w.

Early Warning of Meteorological Drought in Morocco's Atlas Mountains: A Satellite-Augmented Spatiotemporal Deep Learning Approach

Nacer Aderdour[1*], Hassan Rhinane[1], Henri Rueff[2], Mehdi Maanan[1]

[1]*Department of Geology, Faculty of Science Ain Chock, Hassan II University, Casablanca, Morocco*

[2]*Centre for Development and Environment, University of Bern*

**Corresponding author: nacer.ader1@gmail.com*

Acknowledgments

The authors gratefully acknowledge the research environment and support provided by the Centre for Development and Environment, University of Bern, Switzerland, during the laboratory stay.

Abstract

Meteorological drought monitoring in Morocco's High Atlas and Anti-Atlas Mountains remains reactive, yet few predictive studies have combined satellite observations with spatiotemporal deep learning in mountainous terrain. This study develops a convolutional Gated Recurrent Unit (ConvGRU) framework for spatiotemporal forecasting of the 6-month Standardized Precipitation Index (SPI-6) at 1-to-3-month lead times. The model ingests 18 input channels combining CHIRPS precipitation, ERA5-Land reanalysis fields, satellite-derived NDVI, and static terrain features over a 115×210-pixel domain at 5.5 km resolution, with a strict leakage-free temporal split. At a 1-month lead, the ConvGRU achieves an RMSE of 0.714 (9.5% improvement over damped persistence; Diebold-Mariano $p = 0.041$), while this improvement diminishes and becomes non-significant at 2 and 3-month horizons, establishing a clear predictability boundary. The contribution of dynamic climate and vegetation inputs is fundamentally nonlinear: Ridge regression applied to the same inputs captured only 2.3% improvement. In the test-period April 2024 drought (92% area affected), the ConvGRU detected 81% of the drought extent one month in advance compared to 29% by damped persistence. ConvGRU performed best overall relative to ConvLSTM and pixel-wise LSTM baselines, particularly at 1- and 2-month lead times, confirming that spatial convolution is essential for drought detection at longer lead times. These findings demonstrate actionable 1-

month early warning capability and suggest potential for longer-horizon forecasting in topographically complex mountain regions.

Keywords: *drought forecasting, SPI-6, ConvGRU, deep learning, remote sensing, NDVI, Morocco, Atlas Mountains.*

Introduction

Morocco is among the most drought-vulnerable nations in the Mediterranean basin, with an economy that remains substantially dependent on rainfed agriculture and increasingly limited water resources (Bouchaou et al. 2008; Fniguire et al. 2017; Schilling et al. 2020). The High Atlas and Anti-Atlas mountain ranges serve as critical water towers for southern Morocco: seasonal snowmelt and orographic rainfall feed the river systems and aquifers that sustain downstream agricultural zones and rural livelihoods (Marchane et al. 2017). Under accelerating climate change, drought frequency and severity in this region have increased markedly over recent decades (Driouech et al. 2010; Vicente-Serrano et al. 2014; Tramblay et al. 2020), with severe consequences for food security, water availability, and ecosystem integrity.

Despite the clear need for anticipatory drought management, current drought monitoring in Morocco is predominantly reactive; droughts are observed and quantified after their onset rather than forecast in advance (Bijaber et al. 2018). This reactive paradigm limits the effectiveness of water resource planning, agricultural early warning systems, and policy responses (S. Pulwarty and Sivakumar 2014). Transitioning from drought monitoring to drought forecasting requires models that can capture the complex, spatially heterogeneous interactions between atmospheric forcing, land surface response, and hydrological memory that drive drought evolution in mountainous terrain.

The application of deep learning to geophysical forecasting has expanded rapidly, with spatiotemporal architectures demonstrating skill in tasks such as multi-year ENSO prediction (Ham et al. 2019) and precipitation nowcasting (Shi et al. 2015). In the drought domain specifically, the literature has similarly expanded, yet several important gaps persist.. Most studies operate at the station level or single-pixel scale, predicting drought indices at individual locations without exploiting spatial information (Mishra and Singh 2011; Belayneh et al. 2016; Dikshit et al. 2021; Docheshmeh Gorgij et al. 2022; Márquez-Grajales et al. 2024). Those studies that do employ gridded data typically rely exclusively on climate reanalysis products without integrating satellite-derived land surface observations such as vegetation indices or

thermal imagery (Fung et al. 2020). The few studies combining satellite data with deep learning have not specifically targeted semi-arid mountainous regions, where the elevation gradient creates particularly complex interactions among precipitation, vegetation, and temperature (Kogan 1995; AghaKouchak et al. 2015). This motivates a satellite-augmented approach, in which satellite-derived vegetation and land-surface observations complement reanalysis-based climate predictors to capture the full range of drought-relevant signals across heterogeneous mountain terrain.

Furthermore, a critical methodological question remains unaddressed: to what extent do dynamic climate and vegetation predictors, including satellite-derived NDVI, add predictive value beyond lagged drought state, seasonality, and static terrain information? Addressing this question requires a systematic ablation framework comparing simpler and richer predictor sets under identical training, validation, and testing protocols, an approach that remains uncommon in the published literature. To our knowledge, such a controlled ConvGRU-based spatiotemporal drought forecasting design has not yet been applied to Morocco's Atlas Mountain region.

This study addresses the following question: *To what extent do dynamic climate and vegetation predictors, including satellite-derived NDVI, improve multi-horizon SPI-6 drought forecasting skill beyond lagged drought state, seasonality, and static terrain information over Morocco's Atlas region, and how does that added value vary across 1-, 2-, and 3-month lead times?*

We employ SPI-6 as the drought index, computed from CHIRPS precipitation and calibrated over the period 1981–2014. A ConvGRU architecture is used for spatiotemporal sequence-to-map prediction, trained on 32×32 patches with masked mean squared error (MSE) loss over a region-of-interest (ROI) mask. The model predicts the residual over a damped persistence baseline at forecast horizons H = 1, 2, and 3 months. An ablation experiment quantifies the combined contribution of dynamic predictor variables (reanalysis climate fields and satellite-derived NDVI) by comparing models trained with and without these inputs under identical protocols. Multiple SPI timescales (SPI-3, SPI-6, SPI-12) were evaluated in preliminary experiments, with SPI-6 selected based on an optimal balance between autocorrelation structure and deep learning value-added. Binary drought classification was also tested and rejected due to severe class imbalance and loss of intensity information.

This paper makes the following contributions. First, it presents the first spatiotemporal deep learning drought forecast system specifically designed for the Atlas mountain region, using

high-resolution gridded data covering the full spatial domain rather than isolated point locations. Second, it quantifies the added value of dynamic climate and vegetation inputs through a controlled ablation design, demonstrating a 5-7% nonlinear gain attributable to dynamic climate and vegetation inputs. Third, it shows that ConvGRU can provide a 1-month early warning for severe drought events, detecting 81% of drought area compared to 29% by damped persistence in the test-period case study. Fourth, it demonstrates that the contribution of dynamic climate and vegetation inputs is fundamentally nonlinear: Ridge regression applied to identical features captures only a 2–3% RMSE improvement, establishing that a spatiotemporal architecture is essential to exploit these inputs.

1. Study Area and Data

1.1. Study Area

The study domain encompasses the High Atlas and Anti-Atlas mountain ranges of southern Morocco (Figure 1). The High Atlas contains the highest peaks in North Africa, with Jebel Toubkal exceeding 4,000 m, and forms a major hydrological divide between Atlantic-draining catchments to the north and west and endorheic basins to the south and east (Chaponnière et al. 2008). The Anti-Atlas, located farther south, is lower in elevation (typically 1,500 to 2,500 m) and substantially more arid, supporting vulnerable oasis-based agricultural systems that depend strongly on upstream water availability (Karmaoui 2019). Across the domain, climate conditions range from semi-arid Mediterranean on the northwestern slopes to hyper-arid in the southeastern foothills, with strong elevation-dependent gradients in precipitation, temperature, and vegetation cover (Chaponnière et al. 2008). The study area therefore captures a broad physiographic transition from humid mountain slopes to sparsely vegetated desert margins. Seasonal snowpack in the High Atlas provides critical meltwater for downstream river systems and irrigated perimeters across the Draa, Souss, and Ziz-Gheris basins (Marchane et al. 2017). The region of interest (ROI) comprises the land grid cells within the study domain where all predictor variables and the SPI target are jointly defined; the SRTM elevation layer was used as the spatial reference mask. The final grid dimensions and resolution are reported in (Section 2.2).

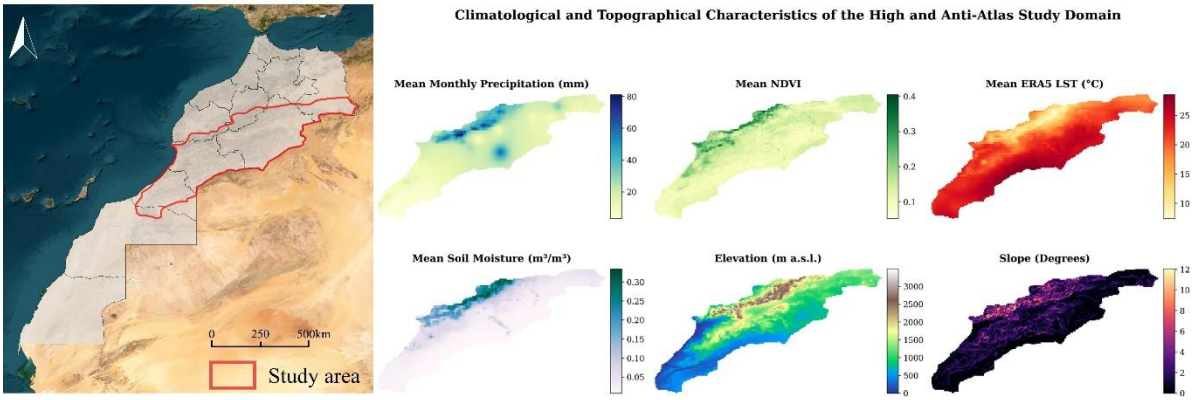


Figure 1. Climatological and topographical characteristics of the High Atlas and Anti-Atlas study domain. Panels show long-term mean values of the six primary input layers: monthly precipitation (mm), Normalized Difference Vegetation Index (NDVI), ERA5-Land skin temperature (degrees C), volumetric soil moisture (m3 m-3), elevation (m a.s.l.), and terrain slope (degrees).

1.2. Input datasets

Monthly input composites were assembled in Google Earth Engine (Gorelick et al. 2017) from satellite and reanalysis sources over the 1981-2026 study period. The extraction pipeline included precipitation from CHIRPS, NDVI from AVHRR and MODIS, three ERA5-Land variables (skin temperature, volumetric soil moisture, and total evaporation), and static terrain layer from SRTM.

Monthly precipitation totals (mm) were derived from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) daily product (Funk et al. 2015), aggregated to monthly sums. CHIRPS provides gridded rainfall estimates from 1981 onward at 0.05° resolution by combining satellite observations with station data.

Three dynamic climate variables were extracted from the ERA5-Land monthly reanalysis dataset (Muñoz-Sabater et al. 2021): volumetric soil water content for the top soil layer (m^3 m^{-3}), total evaporation (mm), and skin temperature (°C). ERA5-Land skin temperature was selected as thermal predictor over satellite-based alternatives such as MODIS land surface temperature, which suffers from cloud-cover gaps in mountain environments and is only available from 2000 onward, whereas ERA5-Land provides continuous coverage from 1981.

Vegetation conditions were represented using the Normalized Difference Vegetation Index (NDVI) at monthly temporal resolution. Two satellite sources were combined: the NOAA

Climate Data Record AVHRR NDVI (V5) for 1981–1999, and MODIS (MOD13A2, Collection 6.1) (Didan 2015) from 2000 onward. To ensure consistency across sensors, a month-wise quantile-mapping procedure was applied to align AVHRR-derived NDVI with the MODIS reference distribution on a per-month basis. Quality flags were applied to both products prior to compositing (Summary QA ≤ 1 for MODIS; range filtering for AVHRR).

Static terrain features, including elevation and slope, were derived from the SRTM digital elevation model at 30 m resolution (Farr et al. 2007) and resampled to the study grid. Latitude and longitude coordinates were also included as static input channels, and a binary ROI mask was appended as the final input channel. All variables were resampled to a common resolution of approximately 5.5 km (EPSG:4326) over the study domain.

In total, the final model input tensor comprised 18 channels: normalized lagged SPI-6 (1), four dynamic climate variables with corresponding validity masks (8), NDVI with its validity mask (2), sinusoidal seasonal encodings of month (2), four static terrain and coordinate features (4), and the ROI mask (1). The four dynamic variables were CHIRPS precipitation, ERA5-Land soil moisture, ERA5-Land skin temperature, and ERA5-Land total evaporation (Figure 2).

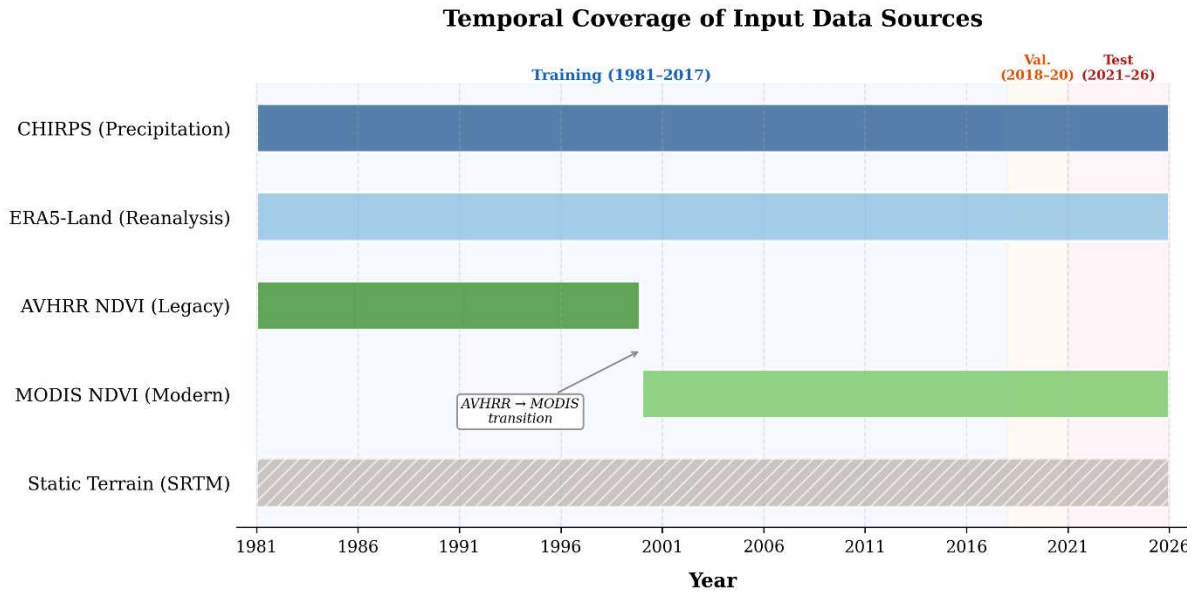


Figure 2. Temporal coverage of the main input data sources used in model development. CHIRPS precipitation and ERA5-Land land-surface variables (soil moisture, evaporation, and skin temperature) are available throughout the 1981–2026 analysis period, whereas NDVI is represented by AVHRR before 2000 and by MODIS thereafter, with month-wise

quantile-mapping harmonization applied across the sensor transition. Static terrain variables (hatched) are time-invariant. Background shading indicates the training, validation, and testing periods.

1.3. SPI-6 computation and climatological analysis

The Standardized Precipitation Index at the 6-month scale (SPI-6) was computed following the methodology of (McKee et al. 1993) as recommended by WMO (Svoboda et al. 2012). Monthly precipitation totals from CHIRPS were accumulated over rolling 6-month windows. For each calendar month, the accumulated precipitation series was fitted to a two-parameter gamma distribution using maximum likelihood estimation, and the fitted cumulative distribution function was transformed to a standard normal variate to yield SPI-6 values. The gamma distribution was calibrated using data from January 1981 to December 2014 (408 months), corresponding to the SPI reference period. This period precedes both the validation and test splits, ensuring that no evaluation-period precipitation influences the drought index computation. Values of SPI-6 below -1.0, -1.5, and -2.0 correspond to moderate, severe, and extreme drought conditions, respectively (McKee et al. 1993; Svoboda et al. 2012).

The marginal distribution of SPI-6 departs slightly from the standard normal, with a modest shift toward negative values consistent with the generally arid character of the domain (Figure 3A). The ROI-mean time series reveals recurring drought episodes throughout the study period, with the embargo window separating the validation and test sets clearly visible (Figure 3B). Autocorrelation remains above 0.5 at a lag of 3 months (Figure 3C), indicating the strong temporal persistence inherent to a 6-month accumulated index; this property is discussed further in (Section 5.1).

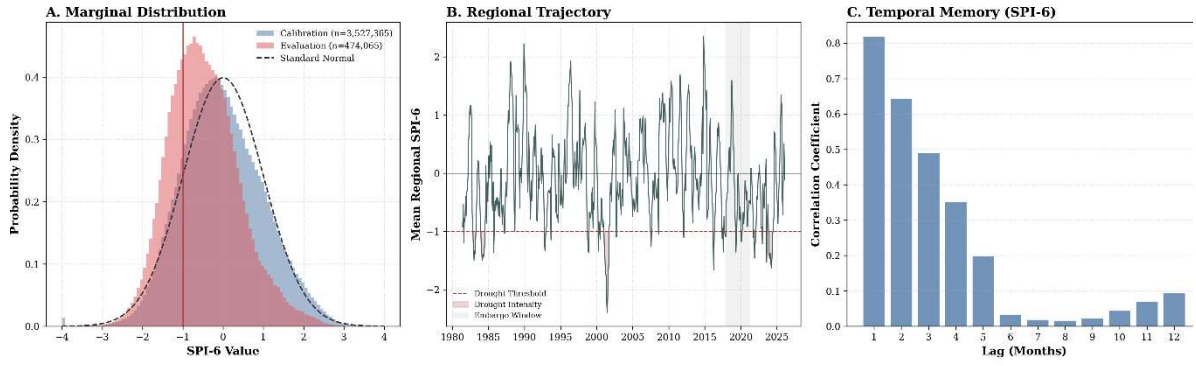


Figure 3. Climatological characteristics of SPI-6 over the study domain. (A) Marginal distribution of pixel-level SPI-6 values for the calibration period (blue, n = 3,527,365) and

evaluation period (red, n = 474,065), with the standard normal distribution shown for reference. The vertical line indicates the moderate drought threshold (SPI = -1.0). (B) ROI-mean SPI-6 time series for 1981-2026, with the drought threshold (dashed red) and the embargo window between validation and test periods (grey shading) indicated. (C) Autocorrelation of the training-period ROI-mean SPI-6 series at lags of 1 to 12 months, illustrating the temporal persistence of SPI-6.

The choice of SPI-6 as the forecast target was informed by preliminary experiments across three accumulation timescales. SPI-3 was evaluated at H = 1 and yielded a ConvGRU RMSE of 1.068, which was worse than Ridge regression (1.057) and only marginally better than damped persistence (1.102), with a correlation of 0.25. The weak autocorrelation of SPI-3 (lag-1 approximately 0.62, lag-3 approximately 0.03) limits its predictability. SPI-12, conversely, is highly persistent (lag-1 approximately 0.97), and Ridge regression already achieved a correlation of 0.82 at H = 1, leaving no margin for deep learning to add value. SPI-6 occupies the productive middle ground: sufficient autocorrelation (lag-1 approximately 0.84, lag-3 approximately 0.50) to support multi-month forecasting, with enough residual complexity that nonlinear models meaningfully outperform linear baselines. Alternative indices such as SPEI (Vicente-Serrano et al. 2010), which incorporate potential evapotranspiration, were considered but not adopted; SPI-6 was retained for computational simplicity, compatibility with the CHIRPS precipitation-only record used here, and consistency with established Moroccan drought monitoring frameworks (Bijaber et al. 2018)

2. Methodology

2.1. Problem formulation

The forecasting task is formulated as a spatiotemporal regression problem. Given a sequence of six consecutive monthly input grids, the model predicts a single two-dimensional SPI-6 map at a specified lead time H months into the future. Each input grid contains 18 channels as described in (Section 2.2). The last month of the input sequence is denoted time t - 1, and the index t is referred to as the anchor month. The prediction target is the SPI-6 field at time t + H, where H takes values of 1, 2, or 3 months. The input sequence therefore spans months t - 6 through t - 1, and no data from the anchor month or later enters the model. Under this convention, the forecast is issued immediately after observing month $t - 1$, and the first unobserved month is the anchor month t.

Rather than predicting the absolute SPI-6 value directly, the model predicts the residual anomaly Δ (delta) relative to a damped-persistence baseline:

$$\Delta(t + H) = SPI_{obs}(t + H) - SPI_{base}(t + H) \quad (1)$$

The final prediction is then reconstructed as:

$$SPI_{pred}(t + H) = SPI_{base}(t + H) + \Delta_{pred}(t + H) \quad (2)$$

where:

$\Delta(t + H)$ is the residual anomaly,

$SPI_{obs}(t + H)$ is observed SPI-6 at the target month,

$SPI_{base}(t + H)$ is the damped-persistence baseline forecast (defined in Section 3.2),

$SPI_{pred}(t + H)$ is the final model prediction,

and $\Delta_{pred}(t + H)$ is the model-predicted residual.

This residual formulation ensures that the deep learning component is trained to capture only the signal that exceeds what temporal persistence and climatology already provide, reducing the risk of the model learning persistence-driven structure already captured by the baseline.

2.2. Damped-persistence baseline

The damped-persistence baseline combines the most recent observed SPI-6 value with the long-term monthly climatology (Mishra and Singh 2011):

$$SPI_{base}(t + H) = \alpha \cdot SPI(t - 1) + (1 - \alpha) \cdot C(m) \quad (3)$$

where:

SPI $(t - 1)$ is the observed SPI-6 at the last month of the input sequence.

$C(m)$ is the training-period mean SPI-6 for the target calendar month m,

α is a blending coefficient estimated separately for each horizon.

The coefficient α was fitted on the training set by minimizing the mean squared error between the damped-persistence prediction and the observed SPI-6 across all ROI pixels:

H = 1: α = 0.629

H = 2: α = 0.479

H = 3: α = 0.344

The decreasing α with increasing lead time reflects the diminishing predictive value of the most recent observation as the forecast horizon extends, with the baseline progressively reverting toward climatology. This baseline serves as both the reference for the residual prediction and the primary benchmark against which the ConvGRU model is compared.

2.3. ConvGRU2D architecture

The model receives the 18-channel tensor input described in (Section 2.2). The primary model is built around a custom Convolutional Gated Recurrent Unit (ConvGRU2D) layer, which extends the standard GRU formulation (Cho et al. 2014) by replacing fully connected operations with two-dimensional convolutions, following the convolutional recurrence paradigm introduced by (Shi et al. 2015) for ConvLSTM.

The ConvGRU2D cell maintains a hidden state h_t with the same spatial dimensions as the input and updates it at each time step through two gates (Cho et al. 2014):

$$z_t = \sigma(W_z * [x_t, h_{t-1}]) \quad (4)$$

$$r_t = \sigma(W_r * [x_t, h_{t-1}]) \quad (5)$$

$$\tilde{h}_t = tanh(W_h * [x_t, r_t \odot h_{t-1}]) \quad (6)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (7)$$

Where:

z_t is the update gate, r_t is the reset gate, $\tilde{h}_t$ is the candidate hidden state, ht is the updated hidden state, x_t is the input at time step t, h_{t-1} is the previous hidden state, W_z, W_r, W_h are learnable

convolutional filter weights, $*$ denotes convolution, $\odot$ denotes element-wise multiplication, σ is the sigmoid function, and square brackets denote concatenation along the channel axis.

The full model architecture consists of the following layers applied sequentially:

(1) Layer normalization applied to the input sequence

(2) ConvGRU2D with 16 filters, 3 × 3 kernel, same padding, dropout = 0.3, L2 weight decay = 1×10^{-4}, returning only the final hidden state

(3) Batch normalization

(4) Conv2D with 1 filter, 1×1 kernel, linear activation, L2 weight decay = 1×10^{-4}

The output is a single-channel two-dimensional map representing the predicted delta-SPI. The ConvGRU2D model accepts inputs of variable spatial dimensions, enabling training on 32 × 32 patches and inference on the full study grid without architectural modification. The full ConvGRU2D pipeline is presented in (Figure 4); the architectures of the two comparison models are provided in Supplementary material. All three architectures (ConvGRU2D, ConvLSTM2D, Pixel-LSTM) share identical training protocols: 16 recurrent units, masked MSE loss, Adam optimizer (lr = 2×10^{-4}), and three random seeds.

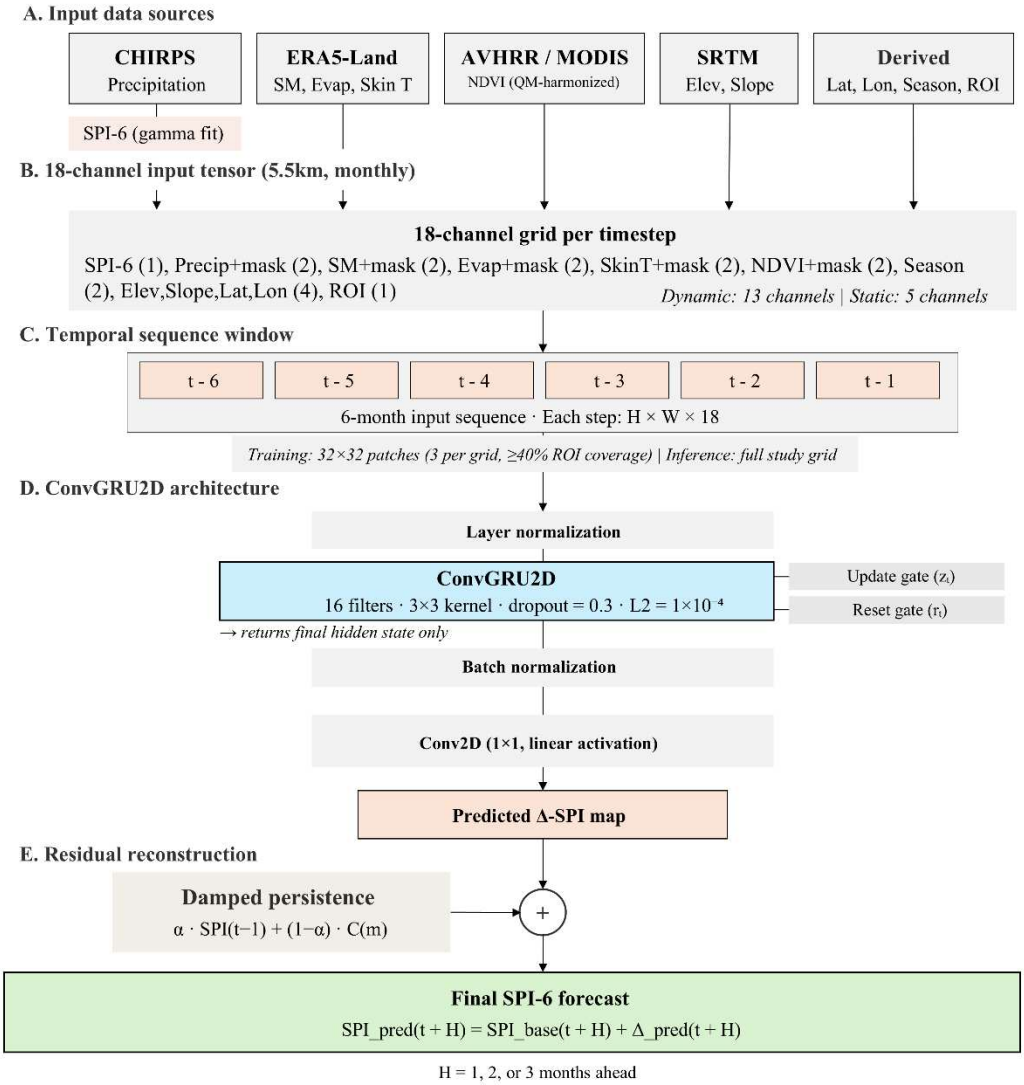


Figure 4. ConvGRU2D pipeline for spatiotemporal SPI-6 drought forecasting. (A) Input data sources resampled to a common 5.5 km grid. (B) The 18-channel input tensor combining dynamic climate variables with validity masks, satellite-derived NDVI, seasonal encodings, and static terrain features. (C) Six-month temporal input sequence, training uses 32 × 32 spatial patches with inference on the full grid. (D) ConvGRU2D architecture: layer normalization, a convolutional GRU cell with update and reset gates (16 filters, 3×3 kernel, dropout = 0.3), batch normalization, and a 1 × 1 convolution producing the predicted delta-SPI map. (E) Residual reconstruction: the predicted anomaly is added to the damped-persistence baseline to yield the final SPI-6 forecast at lead times H = 1,2 or 3 months.

2.4. Comparison architecture

To evaluate the contribution of the gating mechanism and the role of spatial convolution, two alternative models were trained under an identical protocol.

ConvLSTM2D replaces the custom ConvGRU2D cell with TensorFlow's built-in ConvLSTM2D layer, using the same 16 filters, 3 × 3 kernel, and downstream layers (batch normalization followed by Conv2D 1 × 1). The ConvLSTM2D cell differs from the ConvGRU2D in that it maintains both a hidden state and a separate cell state, and uses three gates (input, forget, output) rather than two (update, reset). This comparison provides a controlled assessment of recurrent gating complexity while keeping the spatial convolutional structure broadly comparable. Because TensorFlow's ConvLSTM2D requires fixed spatial dimensions at graph construction time, the model was built with patch dimensions (32 × 32) for training, then rebuilt with full-grid dimensions for test-time inference, with trained weights transferred to rebuilt architecture.

Pixel-LSTM removes spatial convolution entirely. The spatiotemporal input is reshaped so that a standard LSTM with 16 units is applied independently to each pixel's temporal sequence. The LSTM output is passed through batch normalization and a dense layer with a single linear output, then reshaped back to spatial dimensions. This architecture tests whether spatial context contributes to forecast skill or whether per-pixel temporal modeling is sufficient.

All three models share the same number of recurrent units (16), the same downstream projection (batch normalization followed by a single-output linear layer), the same loss function, optimizer, learning rate, training data, and random seeds. The only difference is the recurrent cell and its treatment of spatial dimensions.

2.5. Data splitting and leakage prevention

The dataset was divided into three chronological, non-overlapping periods based on anchor months: training (July 1981 to December 2017; 438 anchor months), validation (January 2018 to December 2020; 36 anchor months), and test (April 2021 to February 2026; 59 anchor months). The earliest valid training anchor is July 1981 because the 6-month input sequence requires data from January 1981 onward.

A gap of approximately 4 months separates the end of the validation period (last validation anchor: December 2020) from the start of the test period (first test anchor: April 2021). Combined with the 6-month SPI accumulation window, this ensures that the precipitation windows contributing to the latest validation targets (ending March 2021 at H = 3) do not overlap with the earliest test targets (beginning May 2021 at H = 1). This embargo was designed to prevent information leakage between the validation and test sets.

Several additional safeguards were implemented to prevent information leakage. The SPI-6 gamma distribution was calibrated on the training period only (1981 to 2014), with no test-period precipitation influencing the probability transformation. All dynamic input variables were normalized using the training-period mean and standard deviation exclusively. Programmatic assertions verified that the training, validation, and test index sets are mutually disjointed.

2.6. Training strategy and protocol

Training was conducted on randomly extracted spatial patches rather than full-grid images to increase the effective number of training samples and reduce GPU memory requirements. For each full-grid sequence, three patches of size 32 × 32 pixels were sampled at random spatial locations, subject to the constraint that at least 40% of the patch area fell within the ROI. Patches with more than 40% missing target values were excluded. This procedure yielded approximately 27,000 training patches across all anchor months. Validation patches were extracted using the same protocol, with independent sampling from the validation period.

At inference time, the full study grid was used without patching. The ConvGRU2D and Pixel-LSTM models accept variable spatial dimensions natively. For ConvLSTM2D, the model was rebuilt at full-grid dimensions with trained weights transferred as described in (Section 3.4).

All models were trained using the Adam optimizer with an initial learning rate of 2 × 10-4 and gradient clipping at a norm of 1.0. The learning rate was reduced by a factor of 0.5 when the validation loss plateaued for 5 consecutive epochs. Training was stopped early if the validation loss did not improve for 20 consecutive epochs, with the best-validation-loss weights restored. The maximum epoch budget was 150. The batch size was 8. The loss function was a masked mean squared error computed only over ROI pixels, ignoring areas outside the study domain.

Each model was trained independently for each combination of horizon (H = 1, 2, 3) and random seed (42, 43, 44), yielding 9 runs per model and 27 runs in total across the three architectures. All runs used the same split logic and preprocessing pipeline.

2.7. Evaluation of metrics and prediction rules

Forecast skill was assessed using both continuous and categorical metrics, computed over all ROI pixels across the full test period.

Continuous metrics included root mean square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (Corr). For the ConvGRU model, 95% bootstrap confidence intervals were computed for all metrics using N = 1,000 temporal block resamples.

Categorical drought detection metrics were computed at three severity thresholds: moderate (SPI < -1.0), severe (SPI < -1.5), and extreme (SPI < -2.0). At each threshold, the probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and F1 score were calculated on a per-pixel basis across the test period.

Statistical significance of forecast improvements was assessed using the Diebold-Mariano (DM) test (Diebold and Mariano 2002), which compares the squared-error loss series of two competing forecasts. A Newey-West heteroscedasticity and autocorrelation consistent (HAC) variance estimator was used to account for serial correlation in the loss differential, and the Harvey-Leybourne-Newbold (HLN) small-sample correction (Harvey et al. 1997) was applied to adjust the test statistic for the limited number of test timesteps (T = 59). The test was applied to evaluate whether the ConvGRU model reduced squared-error loss relative to each baseline. Two prediction rules were applied depending on the output type, as summarized here and applied consistently throughout the Results section.

For scalar metrics, categorical event detection scores, Diebold-Mariano tests, and ROI-mean time-series figures, the ensemble mean prediction across all three seeds was used. This reduces the influence of initialization variability and provides a more stable estimate of expected model performance.

For full-grid spatial maps, permutation feature importance, and the case study illustration, the prediction from the single best-validation-loss seed was used for each horizon (H = 1: seed 43; H = 2: seed 42; H = 3: seed 42). For the future forecast application (Section 4.7), all three seeds were used to produce ensemble-mean predictions with inter-seed uncertainty estimates.

3. Results

3.1. ConvGRU forecast skill.

Table 1 summarizes the test-period performance of the ConvGRU model (ensemble mean across three seeds) alongside the persistence, climatology, damped-persistence, and Ridge regression baselines at all three forecast horizons.

At H = 1, the ConvGRU achieves an RMSE of 0.714 and a Pearson correlation of 0.641, representing a 9.5% RMSE reduction relative to the damped-persistence baseline (RMSE = 0.789). The improvement is substantial compared to Ridge regression, which achieves only a 2.3% RMSE reduction (RMSE = 0.771) using the same 18-channel input features, confirming that the ConvGRU captures nonlinear spatiotemporal structure that a linear model cannot exploit.

At H = 2, skill decreases but the ConvGRU (RMSE = 0.833, Corr = 0.495) still outperforms damped persistence (RMSE = 0.882) by 5.7%, while Ridge achieves 2.7%. At H = 3, the ConvGRU (RMSE = 0.912, Corr = 0.378) offers only a marginal 2.4% improvement over damped persistence (RMSE = 0.934), comparable to the 2.9% achieved by Ridge, indicating that at three months the residual signal exploitable by nonlinear modeling is exhausted.

In a separate ablation experiment conducted under a controlled 50-epoch training budget, the ConvGRU was trained with and without dynamic climate and vegetation inputs using the same three random seeds. Model A (8 channels: SPI history, static terrain, and seasonality only) achieved an RMSE of 0.775 at H = 1, while Model B (18 channels: Model A inputs plus dynamic climate variables and NDVI) achieved an RMSE of 0.725, a relative RMSE reduction of 6.4%. This improvement was consistent across all horizons: 7.1% at H = 2 and 5.4% at H = 3. Model A performed worse than the damped-persistence baseline at H = 2 and H = 3, confirming that the dynamic inputs are essential for forecast skill beyond the shortest lead time. Note that the ablation Model B results differ slightly from the main pipeline results (Table 1) because the ablation used a shorter training schedule to ensure a fair comparison between model variants; the main pipeline results reflect the fully trained model.

Across all horizons, the ConvGRU outperforms simple persistence (no damping) by a wider margin (18.1%, 18.0%, and 17.0% at H = 1, 2, and 3, respectively), while climatology provides no useful forecast (negative correlations at all horizons). The ConvGRU exhibits stable convergence with mild late-epoch validation divergence across all horizons (Figure 5).

Observed versus predicted scatter plots confirm a systematic compression of predictions toward the conditional mean, with linear regression slopes decreasing from 0.52 at H = 1 to 0.25 at H = 3 (**Error! Reference source not found.**), consistent with the MSE training objective discussed in (Section 5.5).

Table 1. *Test-period forecast performance of the ConvGRU ensemble mean and baseline models at three forecast horizons. Bold indicates the best value per horizon. RMSE*

improvement (%) is relative to damped persistence. (Categorical metrics not computed for Ridge regression).

Horizon	***Model***	***RMSE***	***MAE***	***Corr***	***POD***	***FAR***	***CSI***	***RMSE improvement (%)***
H = 1	Persistence	0.872	0.666	0.517	0.578	0.429	0.403	
	Climatology	1.007	0.824	-0.078	0	0	0	
	Damped persistence	0.789	0.613	0.512	0.233	0.242	0.217	
	Ridge	0.771	0.593	0.565				2.3
	ConvGRU	**0.714**	**0.544**	**0.641**	**0.422**	**0.332**	**0.348**	**9.5**
H = 2	Persistence	1.016	0.817	0.352	0.469	0.539	0.303	
	Climatology	1.013	0.83	-0.081	0	0	0	
	Damped persistence	0.882	0.706	0.341	0.067	0.174	0.066	
	Ridge	0.859	0.688	0.439				2.7
	ConvGRU	**0.833**	**0.66**	**0.495**	**0.23**	**0.437**	**0.195**	**5.7**
H = 3	Persistence	1.098	0.89	0.25	0.403	0.598	0.252	
	Climatology	1.02	0.839	-0.072	0	0	0	
	Damped persistence	0.934	0.761	0.231	0.005	0.079	0.005	
	Ridge	0.907	0.741	0.373				2.9
	ConvGRU	**0.912**	**0.733**	**0.378**	**0.136**	**0.508**	**0.119**	**2.4**

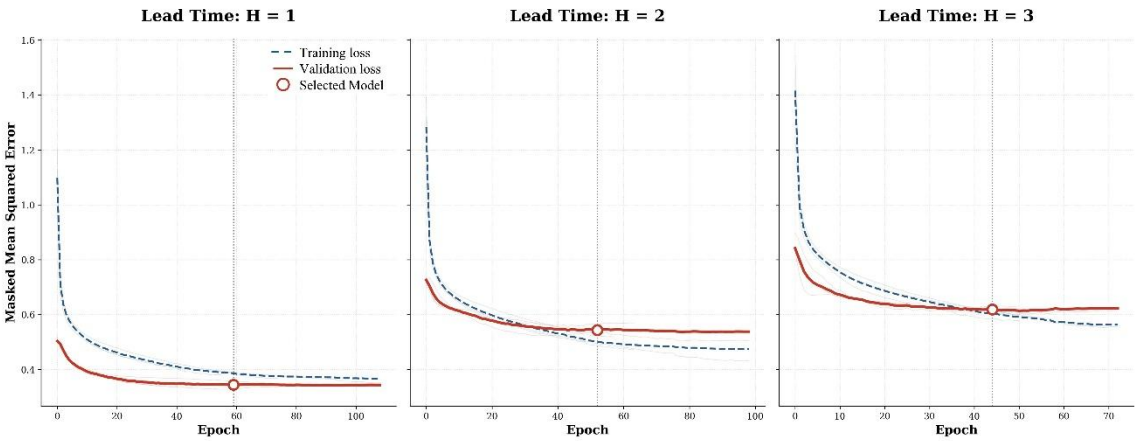


Figure 5. Learning curves of the ConvGRU model for the three forecast horizons. Blue dashed lines show the training loss (masked mean squared error) averaged across three random seeds, and red solid lines show the validation loss. Red circles indicate the epoch at which the best validation loss was achieved and weights were saved. All horizons exhibit stable convergence with mild late-epoch validation divergence, consistent with controlled early stopping.

3.2. Multi-architecture comparison

Table 2 compares the ConvGRU, ConvLSTM, and Pixel-LSTM models across all horizons using the ensemble mean over three random seeds. The ConvGRU achieves the lowest RMSE and highest correlation at H = 1 (RMSE = 0.714, Corr = 0.641) and H = 2 (RMSE = 0.833, Corr = 0.495). At H = 3, all three models converge to similar RMSE values (0.907 to 0.912), though the ConvGRU maintains the highest correlation (0.378 versus 0.355 for ConvLSTM and 0.313 for Pixel-LSTM).

ConvGRU also exhibits the lowest seed-to-seed variability, with RMSE standard deviations of 0.013, 0.006, and 0.010 at H = 1, 2, and 3, compared to 0.017, 0.025, and 0.033 for ConvLSTM. The Pixel-LSTM displays high correlation variance (standard deviation up to 0.088 at H = 3), and 4 of its 9 training runs terminated at epoch 1 without meaningful convergence, indicating training instability when spatial context is removed.

The ConvLSTM showed erratic convergence behavior, with high inter-seed variance particularly at H = 3. Despite having more parameters due to its additional cell state and third gate, it did not outperform the simpler two-gate ConvGRU at any horizon, suggesting that the additional gating complexity does not provide a benefit for this forecasting task (Figure 6).

Table 2. *Multi-architecture comparison on the test set (mean ± standard deviation across three random seeds). Bold indicates the best value per horizon. Categorical metrics are computed at the moderate drought threshold (SPI-6 < -1.0).*

H	*Model*	*RMSE*	*MAE*	*Corr*	*POD*	*CSI*	*F1*
1	**ConvGRU**	**0.714 ± 0.013**	**0.544 ± 0.000**	**0.641 ± 0.004**	**0.422**	**0.348**	**0.517**
	ConvLSTM	0.729 ± 0.017	0.562 ± 0.015	0.632 ± 0.005	0.378	0.323	0.489

	Pixel-LSTM	0.741 ± 0.022	0.572 ± 0.013	0.592 ± 0.056	0.333	0.29	0.448
2	**ConvGRU**	**0.833 ± 0.006**	**0.660 ± 0.000**	**0.495 ± 0.019**	**0.23**	**0.195**	**0.326**
	ConvLSTM	0.849 ± 0.025	0.673 ± 0.020	0.474 ± 0.045	0.222	0.195	0.326
	Pixel-LSTM	0.836 ± 0.009	0.666 ± 0.005	0.453 ± 0.065	0.133	0.125	0.222
3	**ConvGRU**	**0.912 ± 0.010**	**0.733 ± 0.000**	**0.378 ± 0.010**	**0.136**	**0.119**	**0.213**
	ConvLSTM	0.912 ± 0.034	0.737 ± 0.026	0.355 ± 0.042	0.08	0.074	0.135
	Pixel-LSTM	0.907 ± 0.026	0.735 ± 0.023	0.313 ± 0.088	0.026	0.025	0.05

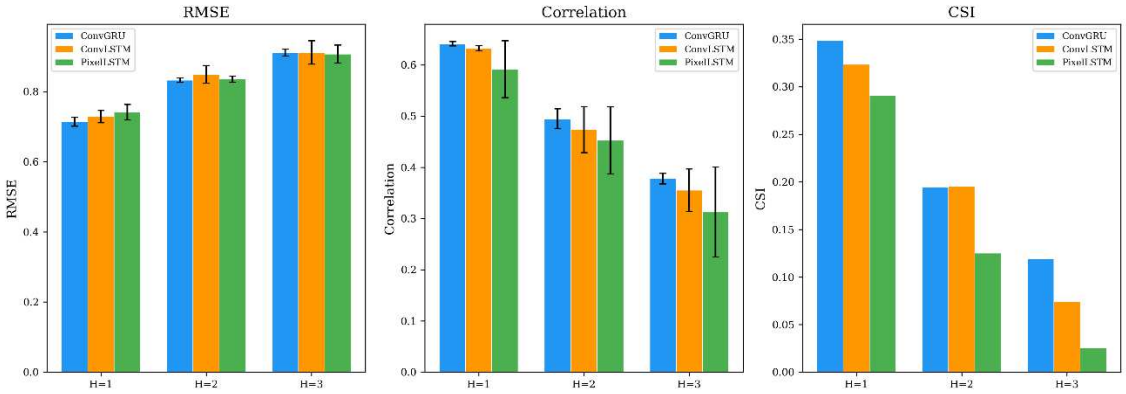


Figure 6. Multi-architecture comparison. RMSE, Pearson correlation, and CSI (moderate drought threshold, SPI < -1.0) for the three architectures across forecast horizons H = 1, 2, and 3. Bars show the mean across three random seeds; error bars indicate one standard deviation.

3.3. Categorical drought detection

At the extreme drought threshold (SPI< -2.0), detection skill is poor for all models, reflecting the severe class imbalance inherent to extreme events. This limitation is discussed further in (Section 5.5).

The advantage of the ConvGRU in drought detection becomes more pronounced across architectures. At H = 1, the ConvGRU CSI (0.348) exceeds that of ConvLSTM (0.323) and Pixel-LSTM (0.290). This gap widens with lead time: at H = 3, the ConvGRU maintains a CSI of 0.119, compared to 0.074 for ConvLSTM and 0.025 for Pixel-LSTM. The near-complete collapse of Pixel-LSTM drought detection at longer horizons (CSI = 0.025 at H = 3) underscores the importance of spatial context for capturing the coherent spatial patterns of drought evolution.

At the extreme drought threshold (SPI < -2.0), detection skill is poor for all models, reflecting the severe class imbalance inherent to extreme events. This limitation is discussed further in (Section 4.5).

3.4. Statistical significance

The Diebold-Mariano test was applied to compare the squared-error loss of the ConvGRU ensemble mean against both the damped-persistence and climatology baselines (Table 3).

Against damped persistence, the ConvGRU improvement is statistically significant at H = 1 (DM = +2.04, p = 0.041), but not at H = 2 (DM = +1.01, p = 0.311) or H = 3 (DM = +0.31, p = 0.758). Against climatology, the improvement is significant at H = 1 (p < 0.001) and H = 2 (p = 0.019), but not at H = 3 (p = 0.156).

These results establish a clear predictability boundary: the ConvGRU provides statistically significant added value over damped persistence only at the 1-month horizon. At H = 2, the model improves upon the baseline numerically, but the improvement cannot be distinguished from sampling variability. At H = 3, the deep learning model offers no statistically demonstrable advantage over the simple baseline, consistent with the diminishing residual signal available beyond the persistence-driven component of SPI-6.

Table 3. *Diebold-Mariano test results comparing the ConvGRU ensemble mean against baseline forecasts. Bold indicates significant results (p < 0.05).*

H	*Comparison*	*DM statistic*	*p-value*	*Significant*
1	**ConvGRU vs. Damped persistence**	**2.04**	**0.041**	**Yes**
1	**ConvGRU vs. Climatology**	**4.53**	**<0.001**	**Yes**

2	ConvGRU vs. Damped persistence	1.01	0.311	No
2	**ConvGRU vs. Climatology**	**2.34**	**0.019**	**Yes**
3	ConvGRU vs. Damped persistence	0.31	0.758	No
3	ConvGRU vs. Climatology	1.42	0.156	No

3.5. Case study: drought detection in validation and test periods

To demonstrate the spatial forecasting capacity of the ConvGRU on fully out-of-sample data, we examine the April 2024 drought, the most severe month in the test period (ROI-mean SPI6 = -1.63, with 92.3% of the domain below the moderate drought threshold). All three forecast horizons target the same event: H = 1 is issued from March 2024 data, H = 2 from February 2024, and H = 3 from January 2024, allowing direct comparison of how skill degrades with increasing lead time for the same drought. At H =1, the ConvGRU achieves an RMSE of 0.554 with a correlation of 0.53, while damped persistence yields an RMSE of 0.824. At H = 2, the ConvGRU retains substantial skill (RMSE = 0.620, r = 0.39), while at H = 3, performance degrades (RMSE = 0.887, r = 0.43) (Figure 7), consistent with the diminishing predictability documented in Table 1.

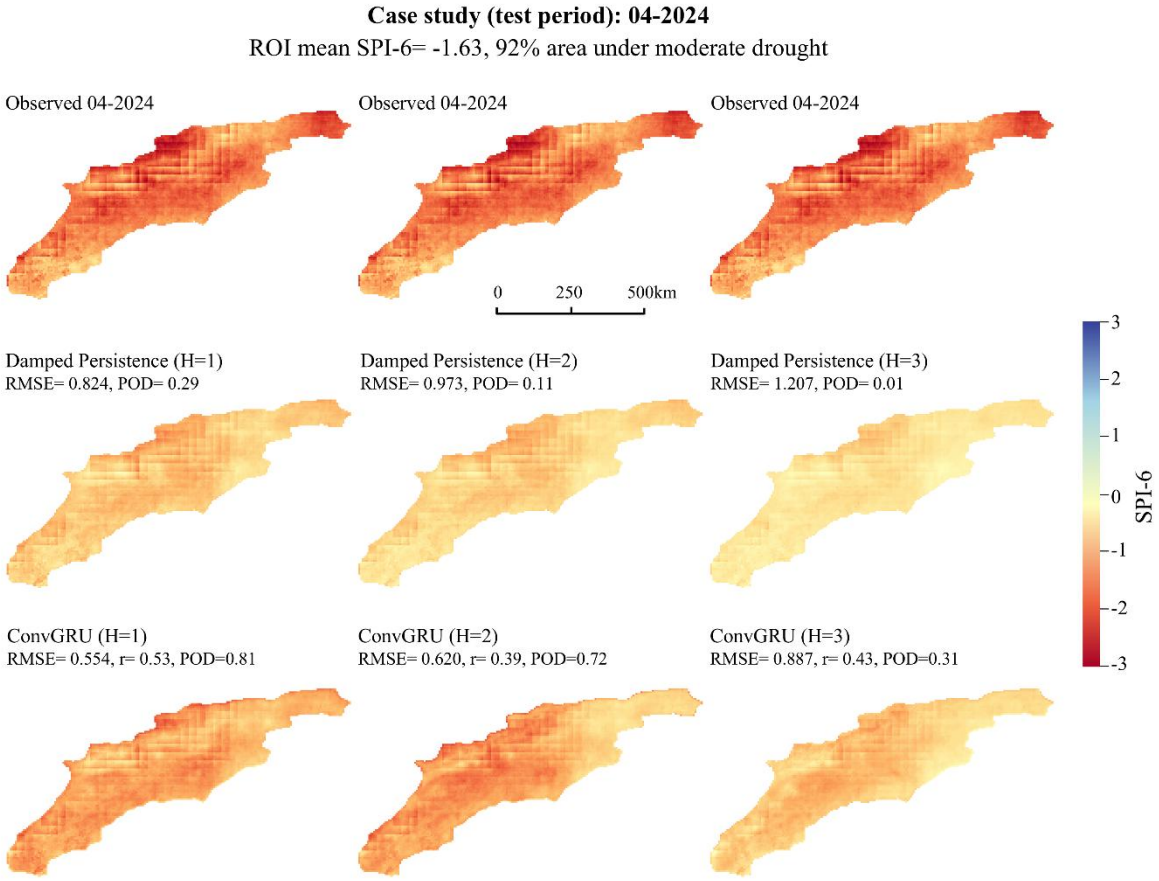


Figure 7. Test-period case study: April 2024 drought event (ROI-mean SPI-6 = - 1.63, 92% area under moderate drought). Continuous SPI-6 fields for three forecast horizons targeting the same event: observed (top row), damped-persistence predictions (Middle row), and ConvGRU predictions (bottom row). Per-panel RMSE, correlation, and POD are annotated. All three horizons predict the same target month; H = 1 is issued from March 2024 data, H = 2 from February 2024, and H = 3 from January 2024.

The binary drought detection maps reveal the contrast more clearly. At H =1, the ConvGRU detects 80.8% of the drought-affected area (POD = 0.808, CSI = 0.776), while damped persistence detects only 29.4%. At H = 2, the ConvGRU retains substantial detection (POD = 0.715, CSI = 0.683), compared to 11.2% for damped persistence. At H = 3, detection skill decreases markedly for the ConvGRU (POD = 0.299) while damped persistence detects less than 1% of the drought extent (Figure 8). A similar detection capability was observed during the May 2019 drought in the validation period, where the ConvGRU detected 90.0% of the drought extent at H = 1 compared to less than 2% by damped persistence (Supplementary Figure S3).

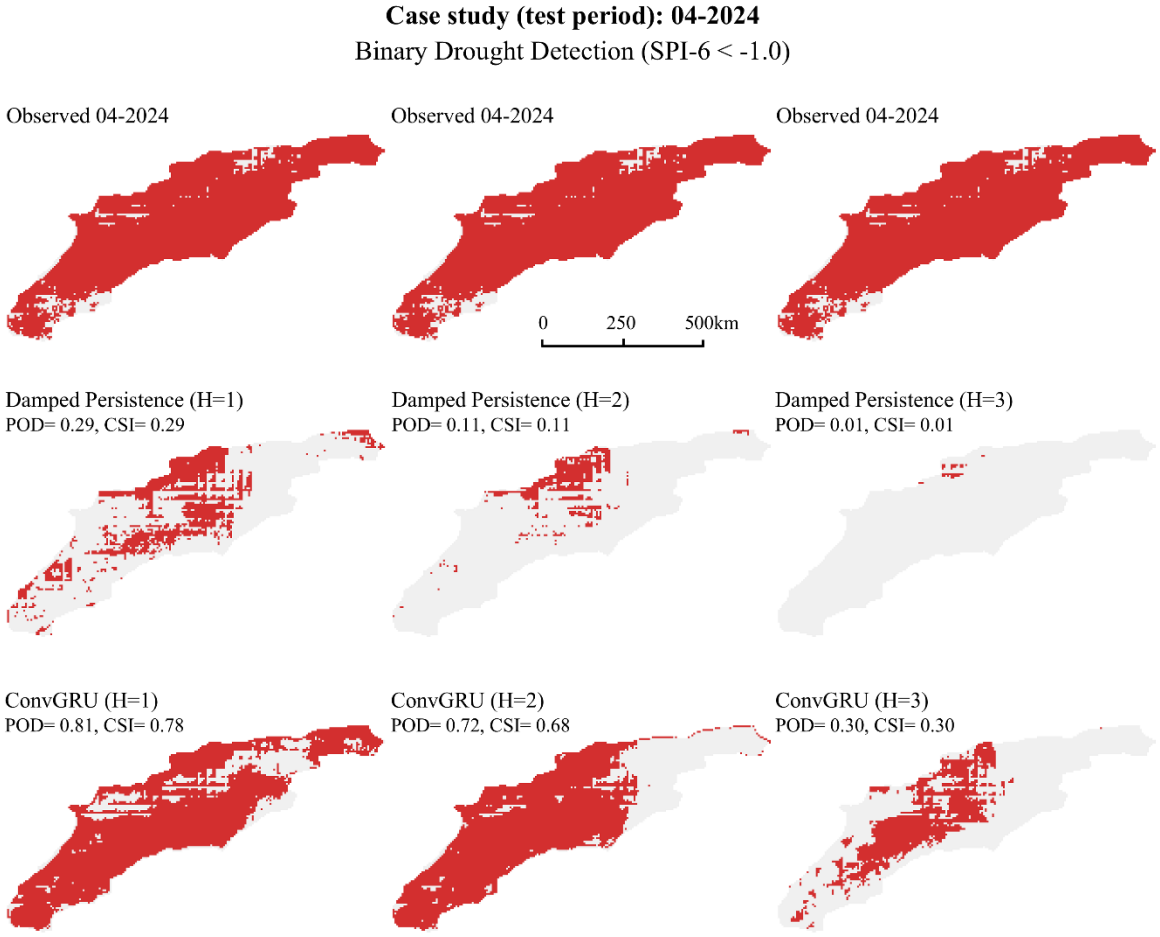


Figure 8. Binary drought detection for the April 2024 case study (SPI-6 < -1.0). Observed drought extent (top row), damped-persistence detection (middle row), and ConvGRU detection (bottom row) at H = 1, 2, and 3. Per-panel POD and CSI are annotated. The ConvGRU detects 81% of the drought area at H = 1 (CSI = 0.78), compared to 29% by damped persistence, confirming that spatial detection skill generalizes to independent test data.

3.6. Spatial forecast skill

Beyond aggregate metrics, the spatial distribution of forecast skill reveals important geographic patterns. At the pixel level, the ConvGRU closely tracks drought onset and recovery at

individual locations, capturing both the timing and magnitude of SPI fluctuations (Figure 9). The damped-persistence baseline, by contrast, consistently underestimates drought severity and lags observed transitions.

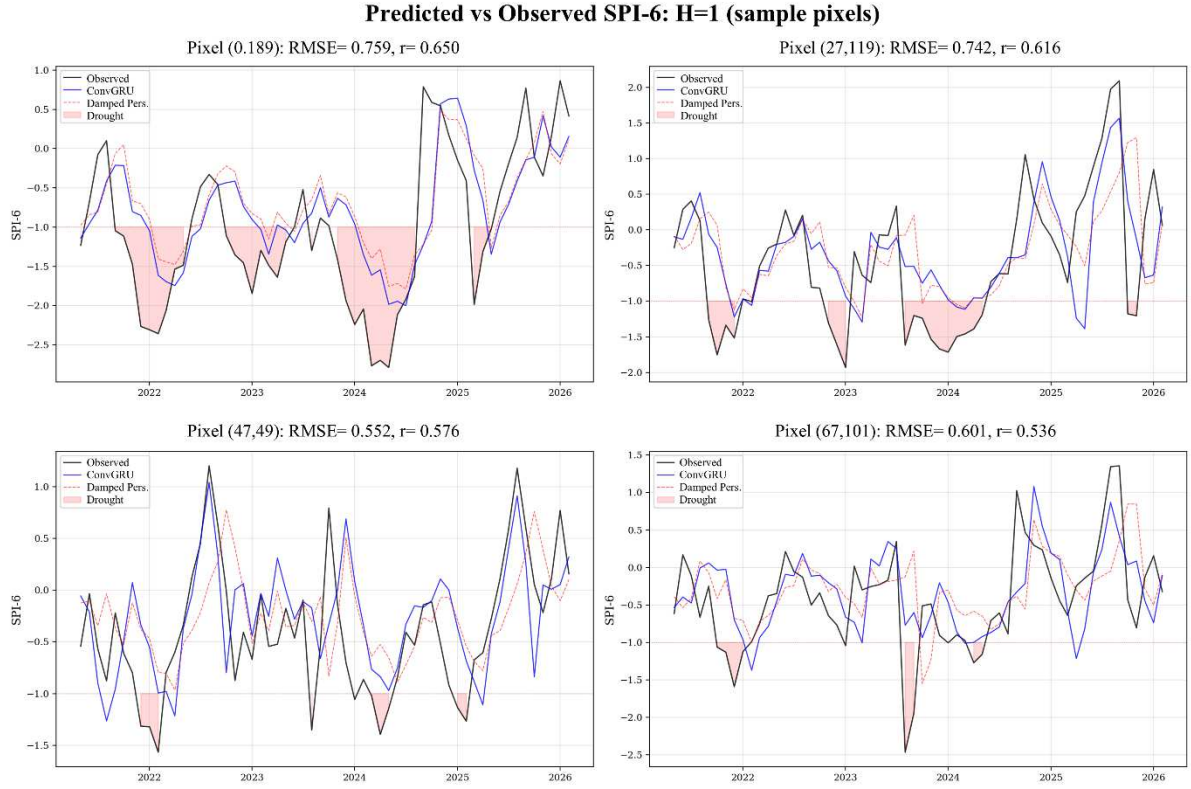


Figure 9. Predicted versus observed SPI-6 time series at four representative pixels for H = 1. Black: observed; blue: ConvGRU; red dashed: damped persistence. Pink shading indicates observed drought periods (SPI-6 < -1.0). Per-pixel RMSE and correlation are shown above each panel.

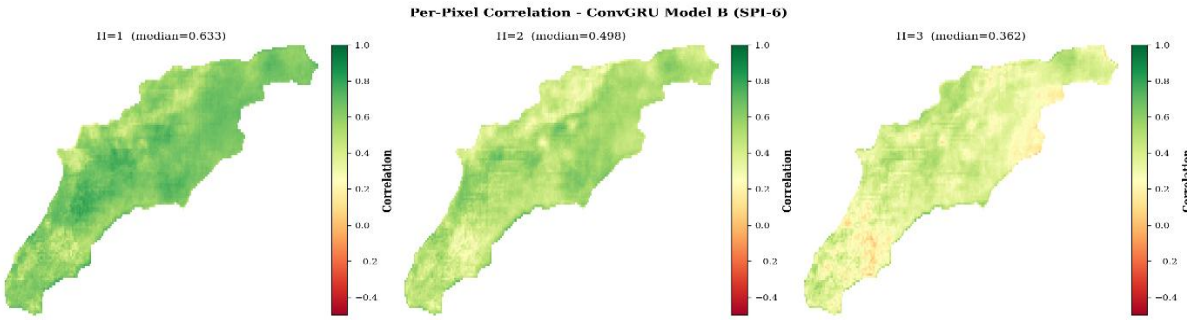


Figure 10. Per-pixel Pearson correlation between ConvGRU predictions and observations for H = 1, 2, and 3. Median correlation values are indicated above each panel. Skill is highest

over the western High Atlas slopes and decreases toward the arid Anti-Atlas margins, with the spatial gradient becoming more pronounced at longer lead times.

462 The spatial distribution of per-pixel Pearson correlation (Figure 10) reveals that skill is highest
463 over the western High Atlas slopes where precipitation variability is strongest. Correlation
464 decreases from a median of 0.63 at H = 1 to 0.49 at H = 2 and 0.36 at H = 3, with the spatial
465 gradient becoming more pronounced at longer lead times. The arid southeastern Anti-Atlas
466 margins show the weakest skill, reflecting the limited dynamic range of SPI-6 in regions with
467 extremely low and irregular precipitation. The ROI-mean SPI-6 time series further confirms
468 that the model closely tracks the observed signal at H = 1, with predictions progressively
469 smoothing and lagging at longer horizons (Figure 11).

470

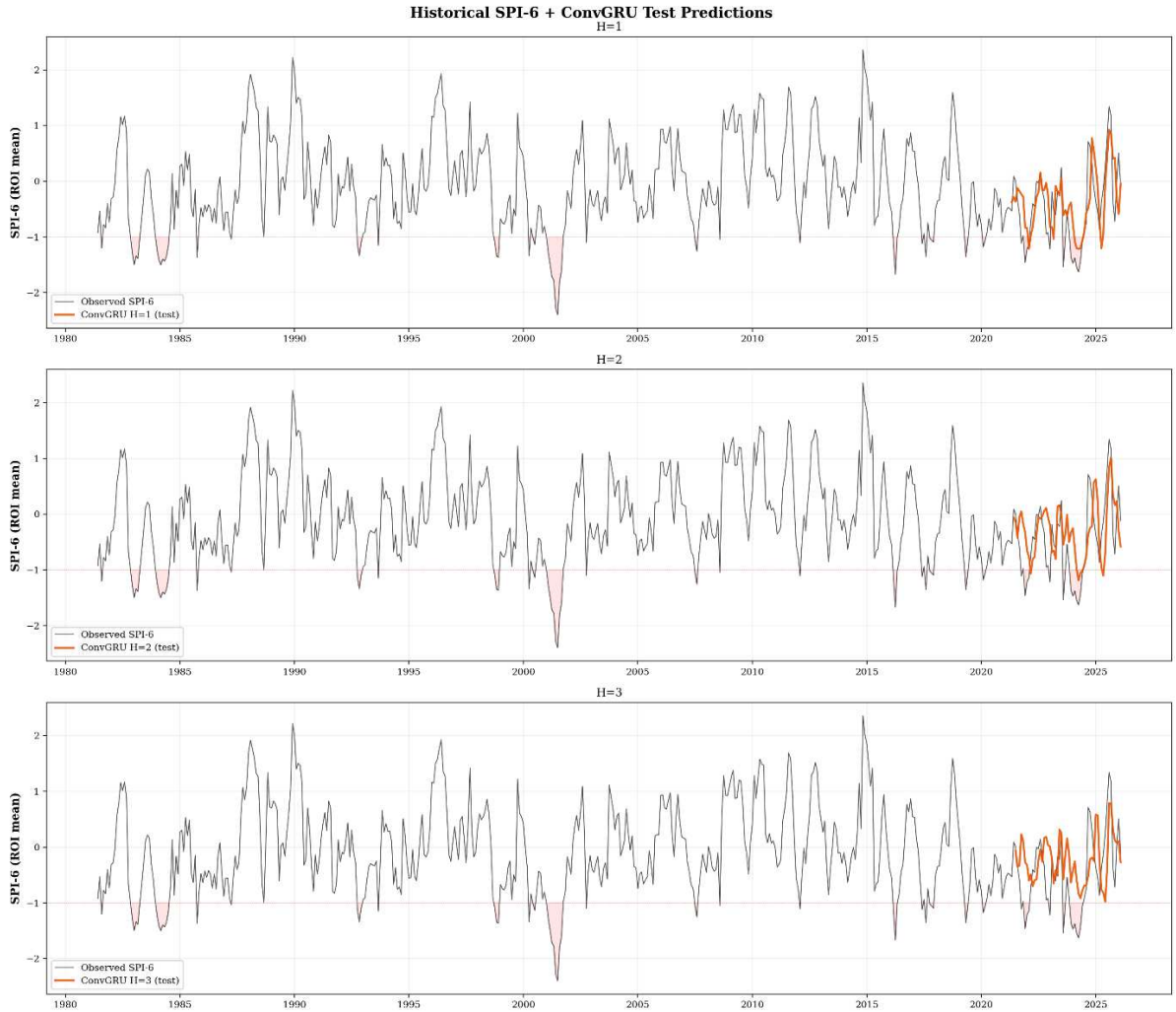


Figure 11. Historical ROI-mean SPI-6 (black) and ConvGRU test-period predictions (orange) for H = 1, 2, and 3. Pink shading indicates observed drought periods (SPI-6 < -

1.0). The model tracks the observed signal most closely at H = 1, with increasing smoothing at longer lead times.

3.7. Ensemble future forecast

To demonstrate operational applicability, the trained ConvGRU was applied beyond the end of the observational record to generate forecasts for March through May 2026 using all three trained seeds. The ROI-mean predictions indicate non-drought conditions across all three horizons (SPI-6 = 0.47 ± 0.25 at H = 1, 0.87 ± 0.44 at H = 2, 0.45 ± 0.43 at H = 3), consistent with the above-average precipitation trend observed in late 2025. The forecast trajectory reflects a marked recovery from the severe drought that affected the region throughout 2023-2024, when ROI-mean SPI-6 values remained below -1.0 for several consecutive months. The transition from drought to above-average conditions aligns with the increased precipitation recorded across the Atlas region during the 2025-2026 wet season (Figure 12). The inter-seed uncertainty increases with lead time, consistent with the diminishing forecast skill established in (Section 4.4).

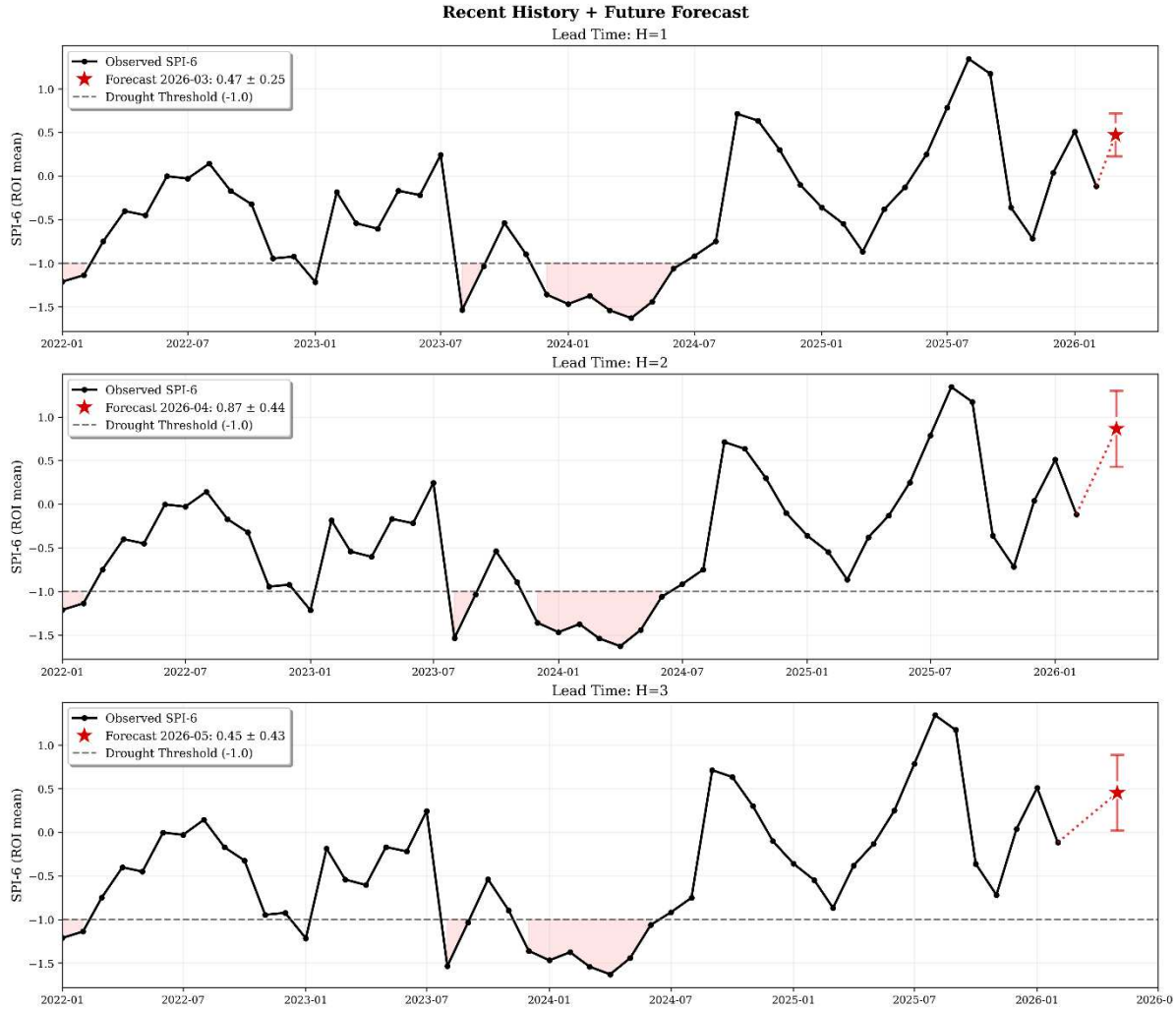


Figure 12. Recent ROI-mean SPI-6 history (black) and ConvGRU future forecasts (red stars) beyond the end of available observations for H = 1, 2, and 3. Error bars indicate inter-seed standard deviation across three trained models (seeds 42, 43, 44. Pink shading indicates observed drought periods.

The corresponding spatial forecast maps confirm predominantly wet conditions across the study domain for all three forecast months, with no pixels exceeding the moderate drought threshold at any horizon (Figure 13). The Western High Atlas shows the strongest positive SPI-6 values, while some drier patches persist along the southeastern Anti-Atlas margins, consistent with the spatial skill patterns observed in the test period. The growing ensemble spread at longer horizons reinforces that operational deployment would require larger ensembles with calibrated confidence intervals.

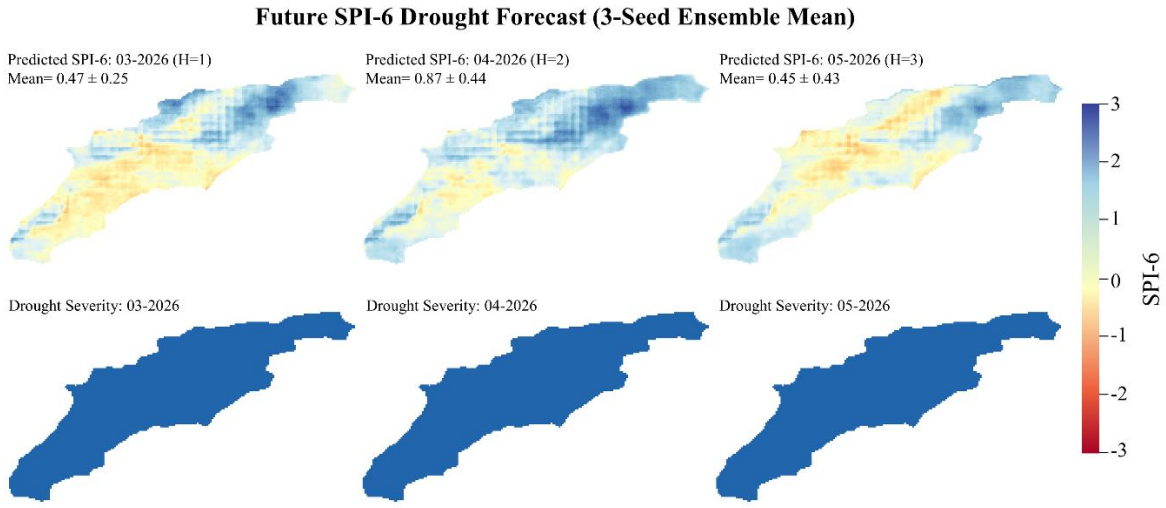


Figure 13. Ensemble-mean future SPI-6 drought forecast beyond available data. Top row: predicted SPI-6 fields for March, April, and May 2026 at H = 1, 2, and 3. With ROI-mean ± inter-seed standard deviation annotated. Bottom row: corresponding drought severity classification. No pixels exceed the moderate drought threshold at any horizon. All predictions are based on the three-seed ensemble mean.

4. Discussion

4.1. SPI-6 autocorrelation and the persistence baseline

A potential concern with SPI-6 forecasting is that the 6-month accumulation window creates substantial overlap between consecutive months: at H = 1, five of the six precipitation months contributing to the target SPI-6 value are already known at forecast time, yielding an inherent autocorrelation of approximately 0.67 at lag 1 (Figure 3C). At H = 3, only three of the six months overlap, reducing the autocorrelation to approximately 0.33. This built-in persistence raises the question of whether the ConvGRU is genuinely learning meaningful atmospheric or land-surface signals or simply exploiting the mechanical overlap.

The residual prediction framework (Section 3.1) was specifically designed to address this concern. By predicting the anomaly relative to a damped-persistence baseline that already captures the persistence-driven component, the ConvGRU is forced to learn only the signal that exceeds what temporal overlap and climatology provide. The damped-persistence baseline itself performs well: at H = 1, it achieves an RMSE of 0.789, which is lower than simple persistence (0.872) and lower than climatology (1.007). The fact that the ConvGRU improves upon this already-strong baseline by a further 9.5% at H = 1, and that this improvement is

statistically significant (DM p = 0.041), indicates that the model captures genuine predictive information beyond the mechanical persistence of SPI-6.

The diminishing improvement at longer horizons (5.7% at H = 2, 2.4% at H = 3) is consistent with this interpretation: as the overlap decreases and the persistence signal weakens, the residual available for deep learning to exploit shrinks correspondingly. The convergence of ConvGRU and Ridge regression performance at H = 3 (2.4% versus 2.9%) further supports this view, suggesting that at three months the exploitable signal is linear and the nonlinear capacity of ConvGRU provides no additional benefit.

4.2. ConvGRU versus ConvLSTM: gating complexity

The ConvGRU consistently outperformed ConvLSTM across all metrics and horizons despite having fewer parameters. The ConvGRU uses two gates (update, reset) compared to the ConvLSTM's three gates (input, forget, output) and an additional cell state, resulting in approximately 33% fewer learnable parameters per recurrent cell. This finding aligns with the broader observation in the sequence modeling literature that the simpler GRU architecture often matches or exceeds LSTM performance on tasks with moderate sequence lengths (Cho et al. 2014; Zhou et al. 2024).

Two factors explain the ConvGRU's advantage in this specific application. First, the input sequences are short (6-time steps), limiting the benefit of the ConvLSTM's cell state, which is primarily designed to preserve information over long sequences. Second, the ConvGRU exhibited markedly more stable training: its seed-to-seed RMSE standard deviation was 0.013 at H = 1, compared to 0.017 for the ConvLSTM and 0.022 for the Pixel-LSTM. The ConvLSTM's higher variance (reaching 0.034 at H = 3) suggests that the additional gating complexity introduces optimization difficulty without providing compensating representational benefit for this forecasting task.

4.3. Spatial convolution versus pixel-wise modeling

The comparison between the spatially convolutional models (ConvGRU, ConvLSTM) and the pixel-wise Pixel-LSTM provides direct evidence that spatial context is essential for drought forecasting. While RMSE differences between models are modest at H = 1 (0.714 versus 0.741), the divergence in categorical drought detection is striking: the ConvGRU CSI of 0.348 is 20% higher than Pixel-LSTM's 0.290, and this gap widens to fivefold at H = 3 (0.119 versus 0.025). The near-complete collapse of Pixel-LSTM drought detection at longer horizons

indicates that pixel-level temporal patterns alone are insufficient to capture the spatially coherent evolution of drought.

This result is physically intuitive. Drought is a spatially extensive phenomenon driven by large-scale atmospheric circulation patterns, and its evolution at any given pixel depends on conditions in neighboring regions. Convolutional filters allow the model to learn spatial covariance patterns (e.g., orographic precipitation gradients, elevation-dependent temperature effects) that a pixel-wise model cannot access. Permutation feature importance analysis confirms that precipitation and seasonality dominate at H = 1, while at longer lead times the relative importance of these features decays and satellite-derived NDVI becomes increasingly relevant (Figure 14), suggesting that the model progressively relies on slower land-surface response signals as the direct atmospheric predictability diminishes.

The cross-horizon importance trend (Figure 14) reveals that precipitation importance decreases from approximately 50% at H = 1 to 38% at H = 3, while NDVI importance increases from near zero to approximately 4%. Although the absolute NDVI contribution remains small, this horizon-dependent behavior suggests that vegetation response acts as a lagged indicator of drought evolution, providing predictive information at lead times where the direct precipitation signal has decayed.

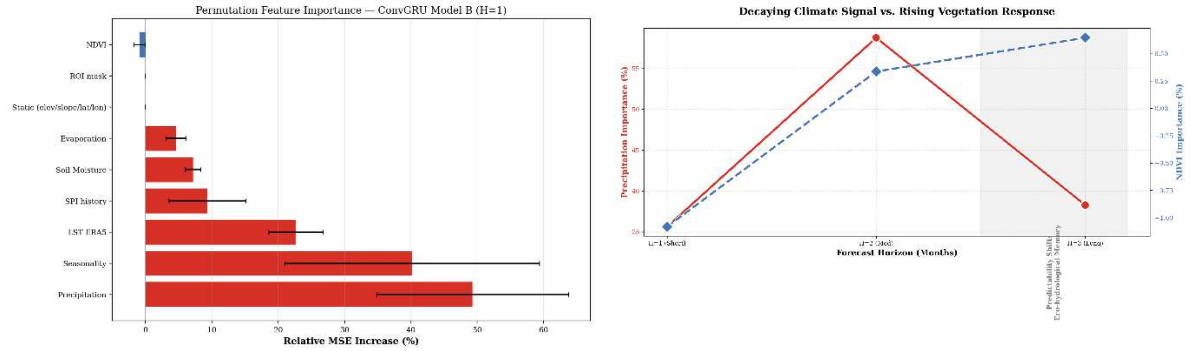


Figure 14. Permutation feature importance and cross-horizon dynamics. (A) Relative MSE increase (%) when each feature group is permuted for the ConvGRU at H = 1. Precipitation and seasonality are the dominant predictors. (B) Precipitation importance (red, left axis) decreases with forecast horizon while NDVI importance (blue dashed, right axis) increases, illustrating the transition from direct atmospheric signals to slower land-surface response at longer lead times.

The spatial distribution of forecast skill (Figure 10) further supports the importance of satellite-derived inputs. Correlation is highest over the western High Atlas slopes, where denser

vegetation cover provides stronger and more variable NDVI and skin temperature signals for the model to exploit. Conversely, skill degrades toward the arid southeastern Anti-Atlas, where sparse vegetation limits the dynamic range of satellite features. This geographic pattern suggests that the ConvGRU's advantage over statistical baselines is concentrated in regions where land surface observations carry the most informational content, and that operational deployment should incorporate pixel-level reliability maps to guide user interpretation.

4.4. Nonlinear contribution of dynamic inputs

The ablation experiment (Section 4.1) revealed that removing dynamic climate and vegetation inputs reduced Model B's 6.4% RMSE improvement over damped persistence at H = 1 to just 1.8% for Model A, and that Model A performed worse than the damped persistence baseline at H = 2 and H = 3. This confirms that SPI history and static terrain alone are insufficient for useful multi-month forecasting. Critically, Ridge regression applied to the same 18-channel feature set captured only 2–3% improvement at all horizons, compared to 6–9% for the ConvGRU. The gap between these two models, which share identical inputs but differ in architecture, indicates that the predictive value of dynamic climate and vegetation features is encoded in nonlinear spatial interactions (Reichstein et al. 2019) that a per-pixel linear model cannot exploit. Convolutional filters capture compound signals, such as the joint behavior of declining NDVI and rising skin temperature during pre-drought periods, which carry more predictive information than either variable in isolation.

Direct quantitative comparison with existing studies is complicated by differences in target indices, spatial scales, and forecast horizons. Most published deep learning drought forecasting studies operate at the station level or single-pixel scale (Belayneh et al. 2016; Dikshit et al. 2021; Docheshmeh Gorgij et al. 2022), predicting drought indices at individual locations. Whereas the present study predicts over a full 115 × 210 gridded domain. The ConvGRU's RMSE of 0.714 at H = 1 is therefore not directly comparable to single-station RMSE values reported in the literature, as gridded prediction over heterogeneous terrain represents a more challenging task. Among spatiotemporal approaches, (Zhang et al. 2023) applied ConvLSTM to pentad-scale SPEI forecasting over China and achieved RMSE = 0.29 with NSE = 0.92 for 1-5 day lead times, confirming that convolutional recurrence captures spatiotemporal drought dynamics effectively, though at a much shorter timescale than the monthly SPI-6 forecast addressed here. (Márquez-Grajales et al. 2024) reviewed deep learning drought prediction methods and noted that most studies lack systematic ablation or statistical significance testing,

both of which the present study provides. The ConvGRU's 9.5% RMSE improvement over damped persistence at H = 1, together with its ability to detect 81% of drought area in the test-period case study, indicates competitive performance, particularly given the topographic complexity of the Atlas domain and the coarse monthly temporal resolution.

4.5. Limits of extreme drought detection

Despite ConvGRU's overall forecast skill, categorical detection of extreme drought events (SPI-6 < -2.0) remains poor across all models and horizons. This limitation reflects a fundamental challenge: extreme drought pixels constitute less than 5% of the training sample, creating severe class imbalance that the masked MSE loss function is not designed to address. The continuous RMSE metric, which weighs all pixels equally, does not penalize the model for missing rare extreme events that occupy a small fraction of the domain.

The model's tendency to underpredict extreme SPI values (Figure 15) reflects a well-known consequence of MSE-based training: the loss function penalizes large errors symmetrically, encouraging predictions to regress toward the conditional mean rather than capturing distributional tails. This regression-to-the-mean effect is exacerbated for extreme events, which occupy a small fraction of the training distribution. In practice, the model predicts SPI values closer to -1.0 even when the observation exceeds -2.0, resulting in correct drought detection at the moderate threshold but failure at the extreme threshold.

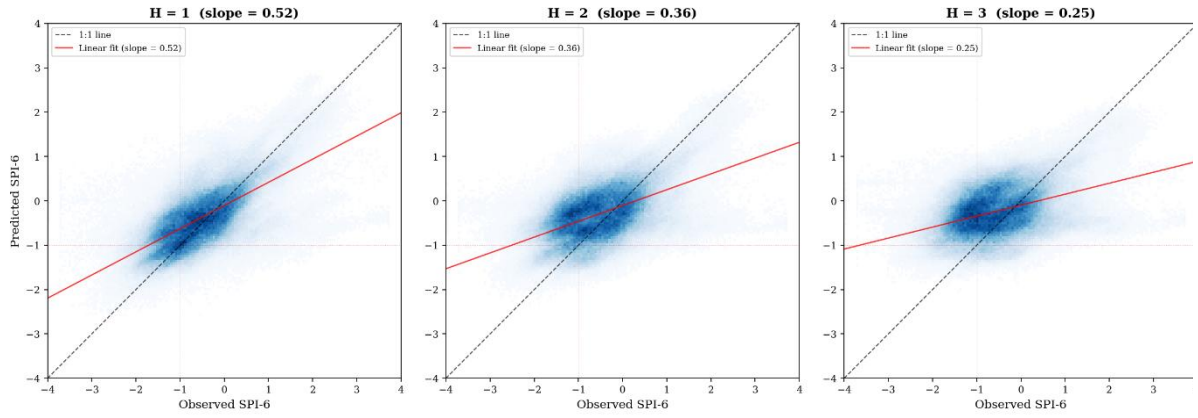


Figure 15. Observed versus predicted SPI-6 for all ROI pixels across the test period (ensemble mean), shown as density scatter plots for H = 1, 2, and 3. The black dashed line indicates the 1:1 reference; the red solid line shows the linear regression fit. Slopes below

1.0 indicate regression toward the conditional mean, with compression increasing at longer lead times. Red dotted lines mark the moderate drought threshold (SPI-6 = -1.0).

Threshold-weighted and focal loss approaches were not explored in this study, as the primary objective was to establish a continuous SPI-6 forecasting baseline; adapting these classification-oriented techniques for spatiotemporal regression remains nontrivial. Addressing this limitation would likely require architectural or training modifications beyond the scope of the current study, such as threshold-weighted loss functions, focal loss adapted for regression (Lin et al. 2017), or a two-stage approach combining continuous SPI prediction with a dedicated extreme-event classifier. These directions represent promising avenues for future work, particularly as climate projections suggest increasing drought severity in the Atlas region (Gumus et al. 2024).

More broadly, the results highlight a predictability boundary for spatiotemporal deep learning applied to SPI-6 forecasting in this region. Statistically significant improvement over damped persistence is achievable only at the 1-month horizon. Beyond this, the persistence-driven component of SPI-6 dominates, and the residual signal available for deep learning becomes indistinguishable from noise. Extending this boundary may require incorporating additional predictive information such as seasonal climate forecasts or soil moisture memory at deeper layers. Large-scale teleconnection indices (e.g., NAO; Knippertz et al., 2003) were tested in preliminary experiments but did not improve forecast skill, suggesting that their influence may already be captured indirectly through the precipitation and temperature fields.

4.6. Limitations

Several limitations of this study should be acknowledged. First, the framework has been evaluated on a single region; its generalizability to other semi-arid mountain environments remains to be established. Second, the test period (April 2021 to February 2026, 59 months) is short, limiting the statistical power of the Diebold-Mariano tests and the number of extreme drought events available for evaluation. Third, the AVHRR-to-MODIS NDVI harmonization relies on quantile mapping, which may not be fully correct for sensor-specific artifacts in mountainous terrain with mixed pixels. Fourth, the residual prediction framework, while effective at preventing the model from learning persistence, may also constrain its ability to predict extreme departures from the baseline. Finally, the patch-based training strategy with 32

× 32-pixel patches may prevent the model from learning large-scale teleconnection patterns that span the full study domain.

5. Conclusion

This study developed a ConvGRU-based spatiotemporal deep learning framework for forecasting SPI-6 drought conditions over Morocco's High Atlas and Anti-Atlas Mountains at 1-to-3-month lead times. The model ingests 18 input channels combining CHIRPS precipitation, ERA5-Land reanalysis fields, satellite-derived NDVI, and static terrain features over a 115 × 210-pixel domain at 5.5 km resolution.

The principal findings are as follows. First, the ConvGRU achieves a statistically significant improvement over the damped-persistence baseline at H = 1 (RMSE = 0.714, 9.5% reduction, DM p = 0.041), confirming that the model captures genuine predictive information beyond the mechanical persistence inherent to SPI-6. Second, the ConvGRU performed best overall relative to ConvLSTM and Pixel-LSTM, particularly at H = 1 and H = 2, while all three models converged at H = 3, demonstrating that a simpler two-gate recurrent architecture with spatial convolutions is sufficient for this task and that spatial context is essential for drought detection. Third, the contribution of dynamic climate and vegetation inputs is fundamentally nonlinear: Ridge regression applied to the same 18-channel inputs captured only 2.3% RMSE improvement at H = 1, compared to 9.5% for the ConvGRU. Fourth, in the test-period April 2024 case study, the model detected 81% of the drought-affected area one month in advance, compared to 29% by damped persistence, demonstrating early warning capability on fully independent data.

The results also delineate clear limits. Statistical significance against the damped-persistence baseline is established only at H = 1; at H = 2 and H = 3, the improvement is not distinguishable from sampling variability. Extreme drought detection remains poor due to class imbalance. The predictability boundary identified here reflects the diminishing residual signal available beyond the persistence-driven component of a 6-month accumulated index.

Several directions merit future investigation. Although large-scale teleconnection indices (e.g., NAO) did not improve skill in preliminary experiments, seasonal climate forecast products or alternative representations of large-scale circulation may offer pathways to extend the predictability horizon. Threshold-adapted loss functions or two-stage architectures may improve extreme event detection. Extending the framework to additional drought indices (e.g., SPEI, soil moisture drought) and to other semi-arid mountain regions would test the

generalizability of the approach. Finally, while a preliminary three-seed ensemble was demonstrated for future forecasting, operational deployment would require larger ensembles with formally calibrated uncertainty bounds.

References

AghaKouchak A, Farahmand A, Melton FS, et al (2015) Remote sensing of drought: Progress, challenges and opportunities. Reviews of Geophysics 53:452–480. https://doi.org/10.1002/2014RG000456

Belayneh A, Adamowski J, Khalil B, Quilty J (2016) Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmospheric Research 172–173:37–47. https://doi.org/10.1016/j.atmosres.2015.12.017

Bijaber N, El Hadani D, Saidi M, et al (2018) Developing a Remotely Sensed Drought Monitoring Indicator for Morocco. Geosciences 8:55. https://doi.org/10.3390/geosciences8020055

Bouchaou L, Michelot JL, Vengosh A, et al (2008) Application of multiple isotopic and geochemical tracers for investigation of recharge, salinization, and residence time of water in the Souss–Massa aquifer, southwest of Morocco. Journal of Hydrology 352:267–287. https://doi.org/10.1016/j.jhydrol.2008.01.022

Chaponnière A, Boulet G, Chehbouni A, Aresmouk M (2008) Understanding hydrological processes with scarce data in a mountain environment. Hydrological Processes 22:1908–1921. https://doi.org/10.1002/hyp.6775

Cho K, Van Merrienboer B, Gulcehre C, et al (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1724–1734

Didan K (2015) MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006

Diebold FX, Mariano RS (2002) Comparing Predictive Accuracy. Journal of Business & Economic Statistics 20:134–144. https://doi.org/10.1198/073500102753410444

Dikshit A, Pradhan B, Huete A (2021) An improved SPEI drought forecasting approach using the long short-term memory neural network. Journal of Environmental Management 283:111979. https://doi.org/10.1016/j.jenvman.2021.111979

Docheshmeh Gorgij A, Alizamir M, Kisi O, Elshafie A (2022) Drought modelling by standard precipitation index (SPI) in a semi-arid climate using deep learning method: long short-term memory. Neural Comput & Applic 34:2425–2442. https://doi.org/10.1007/s00521-021-06505-6

Driouech F, Déqué M, Sánchez-Gómez E (2010) Weather regimes—Moroccan precipitation link in a regional climate change simulation. Global and Planetary Change 72:1–10. https://doi.org/10.1016/j.gloplacha.2010.03.004

Farr TG, Rosen PA, Caro E, et al (2007) The Shuttle Radar Topography Mission. Reviews of Geophysics 45:2005RG000183. https://doi.org/10.1029/2005RG000183

709 Fniguire F, Laftouhi N-E, Saidi ME, et al (2017) Spatial and temporal analysis of the drought
710 vulnerability and risks over eight decades in a semi-arid region (Tensift basin:
711 Morocco). Theor Appl Climatol 130:321–330. https://doi.org/10.1007/s00704-016-
712 1873-z

713 Fung KF, Huang YF, Koo CH, Soh YW (2020) Drought forecasting: A review of modelling
714 approaches 2007–2017. Journal of Water and Climate Change 11:771–799.
715 https://doi.org/10.2166/wcc.2019.236

716 Funk C, Peterson P, Landsfeld M, et al (2015) The climate hazards infrared precipitation with
717 stations—a new environmental record for monitoring extremes. Sci Data 2:150066.
718 https://doi.org/10.1038/sdata.2015.66

719 Gorelick N, Hancher M, Dixon M, et al (2017) Google Earth Engine: Planetary-scale
720 geospatial analysis for everyone. Remote Sensing of Environment 202:18–27.
721 https://doi.org/10.1016/j.rse.2017.06.031

722 Gumus V, El Moçayd N, Seker M, Seaid M (2024) Future projection of droughts in Morocco
723 and potential impact on agriculture. Journal of Environmental Management
724 367:122019. https://doi.org/10.1016/j.jenvman.2024.122019

725 Ham Y-G, Kim J-H, Luo J-J (2019) Deep learning for multi-year ENSO forecasts. Nature
726 573:568–572. https://doi.org/10.1038/s41586-019-1559-7

727 Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared
728 errors. International Journal of Forecasting 13:281–291.
729 https://doi.org/10.1016/S0169-2070(96)00719-4

730 Karmaoui A (2019) Drought and desertification in Moroccan Pre-Sahara, Draa valleys:
731 exploring from the perspective of young people. Geoenviron Disasters 6:2.
732 https://doi.org/10.1186/s40677-019-0118-8

733 Knippertz P, Christoph M, Speth P (2003) Long-term precipitation variability in Morocco and
734 the link to the large-scale circulation in recent and future climates. Meteorology and
735 Atmospheric Physics 83:67–88. https://doi.org/10.1007/s00703-002-0561-y

736 Kogan FN (1995) Application of vegetation index and brightness temperature for drought
737 detection. Advances in Space Research 15:91–100. https://doi.org/10.1016/0273-
738 1177(95)00079-T

739 Lin T-Y, Goyal P, Girshick R, et al (2017) Focal Loss for Dense Object Detection. In: 2017
740 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp 2999–
741 3007

742 Marchane A, Tramblay Y, Hanich L, et al (2017) Climate change impacts on surface water
743 resources in the Rheraya catchment (High Atlas, Morocco). Hydrological Sciences
744 Journal 62:979–995. https://doi.org/10.1080/02626667.2017.1283042

745 Márquez-Grajales A, Villegas-Vega R, Salas-Martínez F, et al (2024) Characterizing drought
746 prediction with deep learning: A literature review. MethodsX 13:102800.
747 https://doi.org/10.1016/j.mex.2024.102800

McKee TB, Doesken NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th Conference on Applied Climatology. American Meteorological Society, Boston, MA, pp 179–184

Mishra AK, Singh VP (2011) Drought modeling – A review. Journal of Hydrology 403:157–175. https://doi.org/10.1016/j.jhydrol.2011.03.049

Muñoz-Sabater J, Dutra E, Agustí-Panareda A, et al (2021) ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst Sci Data 13:4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Reichstein M, Camps-Valls G, Stevens B, et al (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195–204. https://doi.org/10.1038/s41586-019-0912-1

S. Pulwarty R, Sivakumar MVK (2014) Information systems in a changing climate: Early warnings and drought risk management. Weather and Climate Extremes 3:14–21. https://doi.org/10.1016/j.wace.2014.03.005

Schilling J, Hertig E, Tramblay Y, Scheffran J (2020) Climate change vulnerability, water resources and social implications in North Africa. Reg Environ Change 20:15. https://doi.org/10.1007/s10113-020-01597-7

Shi X, Chen Z, Wang H, et al (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting

Svoboda M, Hayes M, Wood DA (2012) Standardized Precipitation Index User Guide. World Meteorological Organization, Geneva

Tramblay Y, Koutroulis A, Samaniego L, et al (2020) Challenges for drought assessment in the Mediterranean region under future climate scenarios. Earth-Science Reviews 210:103348. https://doi.org/10.1016/j.earscirev.2020.103348

Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate 23:1696–1718. https://doi.org/10.1175/2009JCLI2909.1

Vicente-Serrano SM, Lopez-Moreno J-I, Beguería S, et al (2014) Evidence of increasing drought severity caused by temperature rise in southern Europe. Environ Res Lett 9:044001. https://doi.org/10.1088/1748-9326/9/4/044001

Zhang Q, Miao C, Gou J, Zheng H (2023) Spatiotemporal characteristics and forecasting of short-term meteorological drought in China. Journal of Hydrology 624:129924. https://doi.org/10.1016/j.jhydrol.2023.129924

Zhou M, Wu J, Chen M, Han L (2024) Comparative study on the performance of ConvLSTM and ConvGRU in classification problems—taking early warning of short-duration heavy rainfall as an example. Atmospheric and Oceanic Science Letters 17:100494. https://doi.org/10.1016/j.aosl.2024.100494

Statements and Declarations

Data availability

The datasets analyzed in this study are publicly available through Google Earth Engine and the original data providers cited in the manuscript. The derived outputs generated in this study, including processed SPI-6 products, trained model outputs, and figure-generation tables, are available from the corresponding author upon reasonable request.

Code availability

Code availability: The python code for model training, evaluation and figure generation is available from the corresponding author upon reasonable request.

Funding

The authors declare that financial support was received for the research of this article through the project: “Using Vegetation Sensitivity to Climate Change to Predict Agriculture Climate Fragility in the Atlas Mountains Oases of Morocco”. Project number: MS-2024-49. Funded by the HES-SO, Leading House for the Middle East and North Africa.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

NA: Nacer Aderdour; HR: Hassan Rhinane; HRu: Henri Rueff; MM: Mehdi Maanan.

NA: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing-original draft.

HR: Supervision, Validation, Writing-review, and editing.

HRu: Funding acquisition, Writing-review, and editing.

MM: Supervision, Writing-review, and editing.

All authors reviewed the manuscript.

Supplementary Files

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

- SupplementaryMaterial.docx