https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Desert Model Intercomparison Project benchmark framework version 1.0 for assessing land-surface dynamics and surface memory in monthly dust aerosol optical depth over North Africa

Marzieh Mokarram[1], Mohammad Jafar Mokarram[2*], Huayu Lu[3*]

[1]*Department of Geography, Faculty of Economics, Management and Social Sciences, Shiraz University, Shiraz, Iran.*

[2] *School of Electronic Engineering and Intelligent Manufacturing, Anhui Xinhua University, Hefei, China*

[3] *School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China.*

[*] *Corresponding author: Mohammad Jafar Mokarram and Huayu Lu (Email: mjmokarram@axhu.edu.cn, huayulu@nju.edu.cn)*

Abstract

Deserts are the main global source of atmospheric mineral dust, yet large uncertainties remain in the simulation of dust variability across space and time. Part of this uncertainty may reflect the limited representation of dynamic land-surface states and antecedent surface conditions in current dust-model formulations. Here, we develop an interpretable machine-learning assessment framework, here termed the Desert Model Intercomparison Project (DesertMIP) climate–surface–memory benchmark framework version 1.0, to quantify the added value of land-surface dynamics and surface memory for monthly dust aerosol optical depth (AOD) over North Africa during 2003–2020. Three hierarchical configurations were tested: Climate-only, Climate+Surface, and Climate+Surface+Memory. Among the main explanatory models, Random Forest gave the best overall performance. Test-set skill increased from $R^2 = 0.686$ in the Climate-only case to 0.713 after adding surface variables and to 0.736 in the full Climate+Surface+Memory case, while RMSE declined from 0.058 to 0.056 and then to 0.053. The best model also gave MAE = 0.041 and Bias = 0.004. Residuals, defined here as observed minus predicted, were centered close to zero, with a mean of −0.004 and a standard deviation of 0.053, although residual spread increased at higher AOD values. Lag analysis showed a persistent memory-sensitive signal, with the strongest negative associations near a three-month lag for soil moisture ($r \approx -0.75$) and vegetation density (LAI; $r \approx -0.82$), whereas the highest mean cross-validation skill among the tested memory windows occurred at six months. Despite the gains in overall skill, severe dust outbreaks remained difficult to capture. For events above the 95th percentile, the full model gave POD = 0.423, CSI = 0.269, FAR = 0.574, extreme RMSE = 0.170, and extreme Bias = −0.060. These results support a reproducible pilot benchmark structure toward the DesertMIP.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Keywords: Dust AOD; North Africa; Land-surface dynamics; Surface memory; Dust-model benchmarking; Interpretable machine learning; Extreme dust events; DesertMIP.

1 Introduction

Mineral dust is a major component of the Earth's climate system. It affects the radiation budget, cloud processes, nutrient cycles, air quality, and many coupled atmospheric and environmental processes (Kok et al., 2023; Mahowald et al., 2014). Deserts account for about 70–75 % of the global atmospheric mineral aerosol source. Changes in the state and dynamics of desert regions therefore have important consequences at regional and global scales (Kok et al., 2021). Dust also influences climate through processes beyond direct radiative forcing. It modifies cloud behavior, precipitation, and land–atmosphere interactions. These effects generate additional feedbacks in the climate system (Wang et al., 2015).

Earth system models (ESMs) have been the main tools for representing and projecting dust behavior over recent decades. Their development has advanced within major community frameworks such as the Coupled Model Intercomparison Project, and their applications have been central to climate-change assessment (Eyring et al., 2016). However, intercomparison studies still report large uncertainties in dust emission, transport, and deposition (Aryal and Evans, 2021; Huneeus et al., 2011). These uncertainties are especially evident in arid regions such as North Africa and the Middle East, where estimates of dust burden and aerosol optical depth (AOD) differ substantially across models (Pu and Ginoux, 2018). Current model evaluation often relies on bulk metrics such as coefficient of determination (R^2) and root mean square error (RMSE). These metrics summarize overall predictive skill, but they do not by themselves explain why errors occur or where they are concentrated. Recent evidence shows that models may reproduce regional means while still exhibiting strong structural bias under extreme conditions or within specific regions (Adebiyi et al., 2023). A more diagnostic evaluation framework is therefore needed. Such a framework should examine model behavior across surface regimes, error structure, and possible memory effects, rather than relying on domain-wide skill scores alone (Reichstein et al., 2019).

A related gap concerns the relative roles of atmospheric forcing and land-surface state in shaping dust variability and model error. Wind speed is widely recognized as a primary driver of aeolian erosion (Rohrmann et al., 2013). Yet the effects of soil moisture, vegetation cover, and surface roughness remain less systematically quantified, especially within nonlinear diagnostic frameworks (Ginoux et al., 2012; Kim et al., 2024). These surface factors influence dust emission through distinct physical pathways. Roughness reduces near-surface wind stress. Soil moisture increases particle cohesion. Vegetation stabilizes exposed soil and limits sediment entrainment (Zhang et al., 2022). At the same time, many climate and dust-model formulations still depend on simplified or static descriptions of the land surface. Remote-sensing evidence

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

indicates that desert boundaries and land cover are highly dynamic. Seasonal and interannual changes in vegetation, soil moisture, and land use can alter the potential for dust production (Green et al., 2020; Van De Kerchove et al., 2021). Neglect of these dynamics can introduce both spatial and temporal error. It can also reduce the ability of models to reproduce observed patterns of dust activity.

An additional and less explored issue is surface memory. The present land surface does not depend only on instantaneous conditions. It also reflects accumulated environmental history over previous weeks and months. This history may include prolonged drought, gradual vegetation decline, or persistent soil-moisture anomalies (Wen et al., 2025). Such processes can produce delayed effects on sediment availability and erosion sensitivity. These considerations motivate the hypothesis that antecedent surface states may provide additional explanatory power beyond instantaneous surface conditions. However, surface-memory effects have rarely been examined in dust-model diagnostics in a systematic and spatially explicit way. Relationships between model bias, land-cover characteristics, and land-surface dynamics also remain poorly resolved in structural terms (Ridley et al., 2014). A stronger focus on regime-dependent error and lag-sensitive diagnostics may therefore help identify sensitive regions, reveal missing controls, and guide future improvements in dust-model parameterization.

Interpretable machine learning provides a useful way to examine this problem because it can represent nonlinear relationships while still allowing the contribution of different predictor groups to be evaluated. In this study, machine learning is used mainly as a diagnostic tool rather than only as a prediction method. The aim is to test how much additional information is gained when dynamic surface conditions and antecedent surface states are added to climate-only predictors.

The analysis is organized around three hierarchical configurations: Climate-only, Climate plus Surface, and Climate plus Surface plus Memory. This structure allows the study to examine whether land-surface variables improve dust AOD reconstruction, whether antecedent surface conditions provide further information, and whether model errors differ across desert cores, transition zones, and more vegetated environments. Extreme dust events are also evaluated separately, because good mean-state performance does not necessarily imply good performance in the upper tail of the distribution.

The main contribution of this study is therefore not simply the development of another prediction model. Rather, it is the design of a benchmark-oriented diagnostic structure that links climate forcing, land-surface dynamics, surface memory, regime-dependent error, and extreme-event behavior. This structure provides a reproducible pilot framework for future dust-model assessment over dynamic desert environments.

2. Materials and Methods

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

2.1. Overall Study Framework and Methodological Rationale

This study develops a benchmark-oriented framework to evaluate how much information is added by dynamic land-surface conditions and antecedent surface states when reconstructing monthly dust AOD over North Africa. The framework is referred to here as the Desert Model Intercomparison Project (DesertMIP) benchmark framework version 1.0. In this structure, dust AOD is treated as the response of a coupled climate–surface system rather than as the outcome of atmospheric forcing alone.

The relationship was formulated as a nonlinear mapping between climatic drivers, instantaneous surface conditions, memory terms, and dust AOD:

$$Y_t = f(C_t, S_t, M_t) + \varepsilon_t \quad (1)$$

where Y_t denotes dust AOD at time t, C_t denotes climatic and atmospheric predictors, S_t denotes instantaneous land-surface variables, M_t denotes antecedent or memory variables, and ε_t is the residual term.

The unknown function was then approximated using a data-driven nonlinear model (Molnar, 2020; Reichstein et al., 2019):

$$\hat{Y}_t = F_\theta(C_t, S_t, M_t) \quad (2)$$

where F_θ denotes a nonlinear model parameterized by θ. This modelling choice was used because dust response may involve thresholds, nonlinear coupling, and interactions between atmospheric forcing and surface resistance.

To isolate the contribution of each information block, three hierarchical configurations were tested:

$$\hat{Y}_t^{(1)} = F_\theta(C_t) \quad (3)$$

$$\hat{Y}_t^{(2)} = F_\theta(C_t, S_t) \quad (4)$$

$$\hat{Y}_t^{(3)} = F_\theta(C_t, S_t, M_t) \quad (5)$$

The first configuration uses only climatic predictors and serves as the baseline. The second adds dynamic land-surface variables, and the third adds antecedent surface-memory terms. Comparing these three configurations provides a direct way to assess whether surface state and surface memory improve dust AOD reconstruction beyond climate-only forcing.

2.2. Spatial Domain, Temporal Coverage, and Spatio-Temporal Design

The study domain covers North Africa within 20°W–30° E and 10°N–25° N. This region contains some of the world's most important dust-source areas and shows strong gradients in

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

aridity, vegetation, and soil conditions (Kok et al., 2021; Pu and Ginoux, 2018). North Africa is used here as a pilot benchmark domain for DesertMIP because it combines major dust sources with strong ecological transition zones and pronounced land-surface contrasts. These characteristics make it suitable for testing climate-only, surface-aware, and memory-aware assessment designs.

The analysis period extends from 2003 to 2020. Spatial sampling used a regular grid at $0.1° \times 0.1°$resolution. The grid can be expressed as:

$$G = \{(x_i, y_j) \mid x_i \in \lambda,\ y_j \in \phi\} \tag{6}$$

where x_i and y_j denote longitude and latitude coordinates, and λ and ϕ denote the full sets of longitudes and latitudes in the study domain.

Daily data were aggregated to monthly values to reduce short-term noise and to ensure temporal consistency across variables. The monthly value was defined as:

$$X_m = \frac{1}{n}\sum_{t=1}^{n} X_t \tag{7}$$

where X_m is the monthly value, X_t is the daily value, and n is the number of days in the month. Monthly sums were used for cumulative variables such as precipitation. Monthly means were used for continuous variables such as temperature and humidity.

To characterize spatial heterogeneity of the land surface, a composite dryness index was defined from soil moisture and vegetation state:

$$D = 1 - (w_1 \cdot SM + w_2 \cdot NDVI) \tag{8}$$

where D denotes the dryness index, SM denotes soil moisture, $NDVI$ denotes the normalized difference vegetation index, and w_1 and w_2 are weighting coefficients. Spatial transition intensity was then estimated from the gradient magnitude of the dryness index:

$$T = \mid \nabla D \mid \tag{9}$$

where T denotes transition intensity. Larger values of T indicate stronger ecological contrast over space. Based on this indicator, the study area was stratified into three benchmark surface regimes: desert core, transition zone, and vegetated region. This regime-based design forms an important part of the benchmark logic because model skill is expected to differ across these contrasting surface environments.

2.3. Data Sources and Study Variables

2.3.1. Target Variable

The target variable is dust AOD at 550 nm. Data was obtained from the Copernicus Atmosphere Monitoring Service (CAMS). Dust AOD represents vertically integrated atmospheric dust load and is widely used as an indicator of dust activity and dust burden (Kok et al., 2023).

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

A logarithmic transformation was applied to reduce skewness and stabilize model fitting:

$$Y = \log(1 + AOD) \quad (10)$$

where AOD denotes dust AOD and Y denotes the transformed response variable.

Additional climatic and land-surface predictors were obtained from the ERA5 reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Soci et al., 2024). Data was accessed through the Copernicus Climate Data Store (CDS). The selected observational and reanalysis products were chosen not only for the present analysis, but also because they provide a publicly accessible and reproducible basis for future benchmark-oriented experiments within a prototype DesertMIP structure. They also provide consistent spatial and temporal coverage and are compatible with future multi-model comparison workflows.

Quality-control procedures were applied before model development. Unrealistic values were removed. Incomplete records were excluded or flagged during preprocessing. All variables were then harmonized to a common monthly grid.

2.3.2. Climatic and Atmospheric Variables

Climatic predictors were selected to represent the main atmospheric controls on dust emission and transport. Wind speed at 10 m is a primary driver of aeolian erosion (Ginoux et al., 2012). Precipitation influences soil moisture and particle cohesion. Air temperature, relative humidity, boundary layer height, surface pressure, and dew point temperature provide additional information on atmospheric thermodynamic state and near-surface moisture conditions.

Wind speed was calculated from the zonal and meridional wind components:

$$U = \sqrt{u_{10}^2 + v_{10}^2} \quad (11)$$

where $U$is wind speed in m s^{-1}, u_{10} is the zonal wind component, and v_{10} is the meridional wind component.

Wind direction was calculated as:

$$\theta = \mathrm{atan2}(v_{10}, u_{10}) \quad (12)$$

where θ denotes wind direction and $\mathrm{atan2}$ preserves the correct directional quadrant.

Precipitation was expressed in mm or mm $month^{-1}$, depending on the aggregation step. Temperature was expressed in Kelvin or in degrees Celsius, depending on the source product and preprocessing stage. These atmospheric variables define the climate-forcing block in the hierarchical assessment framework.

2.3.3. Land-Surface Variables

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Land-surface predictors were selected to represent the resistance of the surface to wind erosion and its temporal variability. Soil moisture (SM) is a first-order control on sediment cohesion and generally suppresses dust emission as it increases (Kim et al., 2024). The leaf area index (LAI) represents vegetation density and affects roughness, sheltering, and sediment stabilization.

Total vegetation state was represented as:

$$LAI_{\text{total}} = LAI_{hv} + LAI_{lv} \tag{13}$$

where LAI_{hv} and LAI_{lv} denote high- and low-vegetation components, respectively.

Within the benchmark logic of this study, land-surface variables are not treated as secondary descriptors. They form a distinct explanatory block that allows direct comparison between climate-only and climate-plus-surface configurations. This distinction is important for DesertMIP because it makes it possible to test whether model gains arise from atmospheric forcing alone or from improved representation of dynamic surface state.

2.3.4. Memory Variables and Lagged Effects

Environmental memory effects were represented using antecedent windows of land-surface variables. Two related lag treatments were used in this study. First, moving-average memory predictors were constructed for use in the hierarchical explanatory models. Second, a broader lag-correlation analysis was carried out to describe the temporal structure of land-surface associations with dust AOD.

For the explanatory modeling framework, soil-moisture and vegetation memory were defined as moving averages over antecedent months. The baseline memory formulation used in the main Climate+Surface+Memory configuration was based on a 3-month moving average. Soil-moisture memory was defined as

$$SM_{\text{memory},L}(t) = \frac{1}{L}\sum_{i=0}^{L-1} SM(t-i) \tag{14}$$

and vegetation memory was defined as

$$LAI_{\text{memory},L}(t) = \frac{1}{L}\sum_{i=0}^{L-1} LAI(t-i) \tag{15}$$

where L denotes the length of the memory window in months. In the primary explanatory setting, $L = 3$. Additional lag windows of $L = 1$, $L = 3$, and $L = 6$ months were evaluated in a sensitivity analysis to test how reconstruction skill changed with the duration of antecedent surface memory. These lag windows were selected to represent short, intermediate, and seasonally persistent memory scales.

In addition to the predictor-based memory formulations, lag-sensitive correlation diagnostics were computed between dust AOD and key land-surface variables over lags from 0 to 6 months.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

These diagnostics were used to characterize temporal association structure rather than to define the main predictor set. This distinction is important because the lag-correlation analysis serves as a benchmark descriptor of memory-sensitive behaviour, whereas the moving-average memory predictors are part of the explanatory modeling framework.

These memory terms represent delayed land-surface effects that may reflect accumulated drying, vegetation decline, or other forms of surface persistence. Environmental memory effects have been reported to influence land–atmosphere interactions on seasonal timescales (Q. Wen et al., 2025). In the present study, they are treated as observation-constrained descriptors of antecedent surface state rather than as direct proof of causal memory mechanisms.

2.3.5. Dryness Index and Nonlinear Interaction Features

To represent the balance between atmospheric forcing and surface resistance, a process-oriented dryness index was defined as:

$$DI = \frac{U}{SM+\varepsilon} \tag{16}$$

where DI denotes the dryness index, Udenotes wind speed, SM denotes soil moisture, and $\varepsilon$is a small constant that prevents division by zero.

Additional nonlinear interaction terms were also constructed to represent coupled controls between wind forcing and land-surface state. These terms included $SM \times U$and $LAI \times U$. They were introduced to help the model capture situations in which the effect of wind depends on surface moisture or vegetation conditions. This feature construction is consistent with the physical expectation that dust emission is controlled by joint thresholds rather than by single predictors in isolation.

From a benchmark perspective, these interaction variables provide a useful bridge between purely statistical prediction and process-aware interpretation. They also support the broader goal of DesertMIP, which is to assess dust-model behavior under a climate–surface–memory framework rather than under climate forcing alone.

2.4. Variables, Symbols, Units, and Model Roles

Table 1 summarizes the main variable groups, symbols, units, data sources, temporal treatment, and model roles used in this study. The table is structured to distinguish climate drivers, instantaneous surface-state variables, surface-memory terms, process-aware interaction features, regime descriptors, and extreme-event diagnostics. This grouping is also used as the benchmark logic of the prototype DesertMIP framework, because it allows direct comparison between climate-only, surface-aware, and memory-aware assessment tiers.

2.5. Data Preprocessing and Analytical Dataset Construction

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

2.5.1. Spatio-Temporal Harmonization

All variables were harmonized onto a common $0.1^\circ \times 0.1^\circ$ grid and a monthly temporal resolution. This step was required to ensure consistency across products and to allow direct comparison among the three hierarchical predictor configurations. The harmonized value was defined as:

$$X^*_{i,j,m} = A\left(X_t^{(r)}\right) \tag{17}$$

where $X_t^{(r)}$ denotes the original variable from source r at time t, $X^*_{i,j,m}$ denotes the harmonized value at grid cell (i, j) and month m, and Adenotes the harmonization operator. This operator includes spatial remapping and temporal aggregation. The use of a common grid is also important from a benchmark perspective because future DesertMIP-style experiments require transferable inputs and directly comparable outputs across datasets and models.

2.5.2. Quality Control and Missing Data

Quality-control procedures were applied before model development. Values outside physically realistic ranges were removed using variable-specific limits. Missing values were imputed using the median of the training data for each predictor. This choice reduced sensitivity to outliers and avoided leakage of information from the test period into the training workflow. Feature scaling was applied only where required by the modeling algorithm. Tree-based models were trained on the original feature scale, whereas scale-sensitive models were standardized before fitting.

Table 1. Proposed DesertMIP benchmark variable groups

Group	Variable	Role in dust system	Data source	Temporal treatment	Diagnostic use in DesertMIP
Climate drivers	Dust AOD	Target variable for dust variability	CAMS	Daily to monthly; log-transformed for modeling	Benchmark target for all tiers
Climate drivers	u_{10}, v_{10}, U, θ	Wind forcing for erosion and transport	ERA5	Daily to monthly means	Core predictors in DMP-1
Climate drivers	Por TP, T, RH, BLH, SP, DPT	Moisture, thermodynamic state, and boundary-layer controls	ERA5	Monthly means or sums	Core predictors in DMP-1
Instantaneous surface state	SM, LAI_{total}, NDVI	Cohesion, roughness, vegetation shielding, and surface resistance	ERA5 and harmonized land-surface	Monthly means	Added explanatory block in DMP-2

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Group	Variable	Role in dust system	Data source	Temporal treatment	Diagnostic use in DesertMIP
			products		
Surface memory terms	SM_{memory}, LAI_{memory}	Antecedent surface persistence and delayed effects	Derived from monthly surface variables	Moving averages and lag windows	Added explanatory block in DMP-3
Process-aware derived features	DI, $SM \times U$, $LAI \times U$	Coupled forcing–resistance effects and nonlinear thresholds	Derived variables	Monthly	Process-aware interpretation and sensitivity analysis
Regime descriptors	D, T, regime class	Dryness structure and ecological transition intensity	Derived from harmonized surface variables	Spatial diagnostics on monthly grid	Regime-based benchmarking
Extreme-event diagnostics	Q95 threshold, POD, CSI, FAR, extreme RMSE, extreme Bias	Tail behavior and severe-event performance	Derived from observed and predicted AOD	Evaluated on test period	Mandatory tail diagnostics in prototype DesertMIP

2.5.3. Feature Matrix Construction

The analytical feature matrix combined climatic predictors, surface-state variables, memory terms, and derived interaction features. It was defined as:

$$X = [C_t, S_t, M_t, DI, SM \times U, LAI \times U, \ \ldots] \quad (18)$$

The response vector was defined as:

$$Y = [Y_1, Y_2, \ \ldots, Y_n] \quad (19)$$

where C_t denotes climatic predictors, S_t denotes instantaneous surface variables, and M_t denotes memory variables. In the benchmark logic of this study, these feature blocks correspond directly to the progressive information content of the three explanatory tiers. This structure allows the added value of surface and memory information to be quantified in a transparent and transferable way.

For the primary Climate+Surface+Memory configuration, the memory block used the 3-month moving-average formulation, while additional 1-, 3-, and 6-month windows were examined in dedicated sensitivity tests.

2.6. Data Partitioning and Validation Strategy

Temporal ordering was preserved throughout model development in order to reduce information leakage across time. The first 80 % of the time series was used for model training and the remaining 20 % was reserved as an independent temporally out-of-sample test set:

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

$$D_{train} = \{t_1, t_2, \dots, t_k\} \quad (20)$$

$$D_{test} = \{t_{k+1}, \dots, t_n\} \quad (21)$$

Hyperparameter tuning was performed within the training period using blocked time-series cross-validation. This design was chosen to assess temporal generalization while preventing future information from entering model fitting. The present validation should therefore be interpreted as a pilot temporal benchmark rather than a fully independent regional benchmark. Because the split was based primarily on time, reported test performance may still be influenced by spatial autocorrelation among neighboring grid cells. A stricter evaluation of generalization would require spatial blocking or leave-region-out experiments. In a future DesertMIP implementation, these forms of validation should be treated as mandatory components of the protocol.

2.7. Modeling Strategy and Scenario Design

The modeling strategy combined linear and nonlinear algorithms in order to compare simple baselines with more flexible learners. Ridge regression was used as a linear reference model. Random Forest, HistGradientBoosting, and XGBoost were used to represent nonlinear tree-based approaches that can capture threshold behavior and interactions between atmospheric forcing and land-surface state (Reichstein et al., 2019). Algorithm selection was guided by two criteria. The first was the ability to represent nonlinear responses. The second was relative robustness in the presence of correlated predictors.

Three primary explanatory scenarios were evaluated. Scenario 1 corresponds to the Climate-only configuration defined in Eq. (3). Scenario 2 corresponds to the Climate+Surface configuration defined in Eq. (4). Scenario 3 corresponds to the Climate+Surface+Memory configuration defined in Eq. (5). For clarity and future protocol transferability, these three explanatory scenarios are also referred to here as pilot DesertMIP tiers: DMP-1 (climate-only), DMP-2 (climate plus surface), and DMP-3 (climate plus surface plus memory). This tiered structure converts the present analysis into a benchmark architecture rather than a single-model exercise. Within DMP-3, the baseline memory predictors were defined using a 3-month moving window, and alternative 1-, 3-, and 6-month formulations were evaluated separately to assess the sensitivity of benchmark performance to memory-window length.

An ablation analysis was used to quantify the incremental contribution of each predictor block. This analysis tested how model skill changed after removal of surface-state variables, vegetation terms, soil-moisture terms, or memory features. From a DesertMIP perspective, this step provides a diagnostic link between aggregate model performance and the information content of specific variable groups.

In addition to the three explanatory tiers, an auxiliary autoregressive benchmark was also evaluated by adding lagged dust AOD ($dust\_lag1$) as a predictor. This benchmark is referred

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

to here as DMP-AR. It was not treated as part of the primary explanatory comparison because it includes prior information from the target variable itself. It therefore provides a statistical upper-bound reference rather than a directly comparable process-oriented scenario. Its purpose is to show how much predictive skill can be gained when persistence in the target series is used explicitly.

2.8. Model Performance Metrics

Model evaluation was organized around two diagnostic levels. The first level includes **core diagnostics** that quantify overall predictive skill and benchmark performance across regimes and event classes. The second level includes **extended diagnostics** that describe how and where model behavior changes across the predictor hierarchy. This distinction is useful for the present analysis and also provides a transferable metric structure for a prototype DesertMIP benchmark.

The core diagnostics include the coefficient of determination (R^2), root mean square error $(RMSE)$, mean absolute error (MAE), mean bias, regime-specific RMSE, and extreme-event detection metrics. The coefficient of determination was defined as

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{22}$$

The root mean square error was defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \tag{23}$$

The mean absolute error was defined as

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | Y_t - \hat{Y}_t | \tag{24}$$

To ensure a consistent sign convention across the manuscript, mean bias was defined as predicted minus observed:

$$Bias = \frac{1}{n}\sum_{t=1}^{n}(\hat{Y}_t - Y_t) \tag{25}$$

Under this definition, positive bias indicates overestimation and negative bias indicates underestimation.

Extreme dust events were defined using the 95th percentile of the observed AOD distribution in the evaluation dataset. Event-detection skill was assessed using probability of detection (POD),

$$POD = \frac{Hits}{Hits + Misses} \tag{26}$$

critical success index (CSI),

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

$$CSI = \frac{Hits}{Hits+Misses+FalseAlarms} \quad (27)$$

and false alarm ratio (FAR),

$$FAR = \frac{FalseAlarms}{Hits+FalseAlarms} \quad (28)$$

Core diagnostics were computed for each model configuration and for each benchmark surface regime. The extended diagnostics include conditional-bias curves, lag-dependence diagnostics, spatial error maps, and variable-group contribution analysis. Within the prototype DesertMIP logic, the core diagnostics define minimum benchmark outputs, whereas the extended diagnostics provide process-oriented interpretation of performance differences across tiers.

2.9. Explainability Analysis and Variable Contribution

Explainability analysis was used to quantify how predictor importance and response structure changed across the three hierarchical configurations. Global predictor importance was estimated from the fitted model using error-based importance scores, and complementary analyses were used to compare the relative role of climatic, surface, and memory variables (Molnar, 2020). These diagnostics were intended to support interpretation of model behavior rather than to establish direct physical causality.

Partial-dependence analysis was used to summarize average response patterns for selected predictors:

$$PD(X_j) = E_{X_{-j}}[F_\theta(X)] \quad (29)$$

where $PD(X_j)$denotes the partial dependence of predictor X_j, and X_{-j}denotes the set of all remaining predictors.

To examine whether the addition of surface and memory variables changed systematic error, scenario-wise differences in mean bias were defined as

$$\Delta Bias_S = Bias(C) - Bias(C+S) \quad (30)$$

$$\Delta Bias_M = Bias(C+S) - Bias(C+S+M) \quad (31)$$

where C denotes the Climate-only configuration, $C+S$ denotes the Climate+Surface configuration, and $C+S+M$ denotes the Climate+Surface+Memory configuration. Positive values of $\Delta Bias$indicate reduced absolute bias after the addition of the corresponding predictor block.

These diagnostics were examined at the full-domain scale and across benchmark surface regimes. This design allows the analysis to test whether soil moisture, vegetation state, and antecedent memory terms contribute differently across desert cores, transition zones, and more vegetated environments. Within a future DesertMIP framework, these explainability diagnostics are intended as standardized interpretive outputs rather than as stand-alone causal

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

evidence.

2.10. Computational Environment and Reproducibility

All analyses were conducted in the Python scientific-computing environment. Core libraries included xarray, netCDF4, NumPy, pandas, scikit-learn, and XGBoost where required. Reproducibility was supported through fixed random seeds, explicit preprocessing rules, and a documented workflow from data harmonization to model evaluation.

The computational pipeline included data acquisition, spatial and temporal harmonization, quality control, construction of base predictors, generation of interaction and memory features, assembly of the analytical feature matrix, temporal partitioning, model fitting, hyperparameter tuning, benchmark evaluation, and explainability analysis. To support reproducibility beyond the present manuscript, the archived workflow includes the exact repository version, the exact Zenodo DOI, experiment configuration files for all benchmark tiers, region masks for the three benchmark regimes, plotting scripts, and benchmark summary tables. Together, these archived materials allow the main results of the pilot benchmark to be reproduced from the documented inputs and configuration settings. This level of documentation is especially important for a benchmark-oriented study because future DesertMIP-style applications require transferable experiment definitions rather than only narrative method descriptions.

2.11. Prototype DesertMIP Benchmark Logic and Protocol

This study defines a prototype DesertMIP benchmark logic and protocol for transferable dust-model assessment. The proposed framework should be interpreted as a pilot proof of concept rather than as a finalized community intercomparison standard. Its main objective is to diagnose the added value of dynamic land-surface information and antecedent surface memory in dust-model assessment within a structured and reproducible benchmarking architecture.

The pilot benchmark domain is North Africa at monthly resolution over 2003–2020. This domain was selected because it combines strong dust activity with pronounced gradients in aridity, vegetation cover, and ecological transition intensity, making it a suitable test bed for climate–surface–memory benchmarking. The benchmark is organized around three primary explanatory tiers: DMP-1, corresponding to the Climate-only configuration; DMP-2, which adds instantaneous land-surface variables; and DMP-3, which adds antecedent surface-memory terms. In addition, DMP-AR is included as an optional autoregressive reference that provides a statistical upper-bound benchmark but is not treated as part of the main explanatory hierarchy.

Mandatory benchmark outputs include overall skill metrics, regime-specific skill metrics, extreme-event diagnostics, lag-sensitive diagnostics, and spatial bias maps. Benchmark evaluation is performed across three required reporting strata: desert core, transition zone, and vegetated region. These regimes are treated as essential benchmark classes because model

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

behavior is expected to differ substantially across them.

In this sense, the present implementation provides a reproducible pilot benchmark structure for climate–surface–memory assessment of dust models. It does not yet constitute a community-finalized DesertMIP protocol, but it offers a structured blueprint that can be extended in future developments through additional domains, stricter spatial validation strategies, and direct comparison with process-based dust models.

3. Results

3.1. Overall Evaluation of Model Performance in Reconstructing Dust AOD

3.1.1. General performance across the three explanatory configurations

The three explanatory configurations show a clear and ordered increase in reconstruction skill as additional predictor groups are introduced. Across most algorithms, performance improves from the Climate-only configuration to Climate+Surface and then to Climate+Surface+Memory. This pattern indicates that dynamic land-surface information and antecedent memory terms add useful explanatory information beyond climate-only forcing.

Within the Climate-only configuration, XGBoost achieved the highest test performance, with $R^2 = 0.700$, $MAE = 0.043$, and $RMSE = 0.057$. Random Forest followed closely, with $R^2 = 0.686$and $RMSE = 0.058$. After surface variables were added, Random Forest became the best-performing model, with $R^2 = 0.713$, $MAE = 0.042$, and $RMSE = 0.056$. In the full Climate+Surface+Memory configuration, Random Forest again provided the highest overall skill, with $R^2 = 0.736$, $MAE = 0.041$, and $RMSE = 0.053$. This was the best test performance among all explanatory models examined in this study.

The same comparison also shows that tree-based models consistently outperformed the linear Ridge model. This result suggests that the relationship between dust AOD and the predictor set is not well represented by a purely linear structure. Instead, the reconstruction problem appears to benefit from methods that can represent nonlinear responses and interactions.

Although XGBoost slightly outperformed Random Forest in the Climate-only configuration, Random Forest was selected as the reference model for subsequent interpretation. Its performance was strongest in the two surface-aware configurations that are most relevant to the climate–surface–memory framework. It also provided the best overall result in the full explanatory setting. For this reason, Random Forest was used as the main interpretive model in the analyses that follow (Table 2).

Table 2. Performance metrics of machine-learning models for dust AOD reconstruction under three predictor configurations.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Scenario	Model	Train (R^2)	Test (R^2)	Test (MAE)	Test (RMSE)	Bias
Climate-only	XGBoost	0.859	0.7	0.043	0.057	0.007
Climate-only	RandomForest	0.979	0.686	0.044	0.058	0
Climate-only	HistGB	0.948	0.629	0.048	0.063	0.021
Climate-only	Ridge	0.601	0.594	0.052	0.066	0.005
Climate+Surface	RandomForest	0.976	0.713	0.042	0.056	-0.002
Climate+Surface	XGBoost	0.859	0.706	0.043	0.056	-0.001
Climate+Surface	HistGB	0.945	0.675	0.045	0.059	-0.011
Climate+Surface	Ridge	0.624	0.614	0.051	0.064	0.004
Climate+Surface+Memory	RandomForest	0.983	0.736	0.041	0.053	0.004
Climate+Surface+Memory	XGBoost	0.857	0.721	0.042	0.055	0.003
Climate+Surface+Memory	Ridge	0.678	0.676	0.046	0.059	0.003
Climate+Surface+Memory	HistGB	0.951	0.672	0.045	0.059	-0.004

3.1.2. Comparative performance assessment across configurations

The comparison among the three explanatory configurations shows a gradual improvement in model skill as additional information is introduced. For the Random Forest model, test R^2 increased from 0.686 in the Climate-only configuration to 0.713 in Climate+Surface and to 0.736 in Climate+Surface+Memory. Over the same sequence, test RMSE decreased from 0.058 to 0.056 and then to 0.053. Other nonlinear models showed broadly similar behavior, although the size of the improvement differed among algorithms.

This stepwise improvement suggests that land-surface variables add information that is not fully captured by atmospheric predictors alone. The additional gain after including memory terms is smaller, but it still indicates that antecedent surface conditions contain useful information beyond the instantaneous surface state. In the proposed benchmark hierarchy, these results define the basic skill progression from DMP-1 to DMP-3.

3.1.3. Relationship between observed and predicted values

Figure 1b compares observed and predicted dust AOD for the best-performing explanatory model, Random Forest under the Climate+Surface+Memory configuration. Most points are concentrated near the 1:1 line, showing that the model captures much of the observed variability. The agreement is strongest for low and moderate AOD values, where most observations are located.

The spread becomes larger at higher AOD values. Several high observed AOD cases fall below the 1:1 line, indicating underestimation of intense dust conditions. This behavior shows that the model performs well for the dominant range of the data but becomes less reliable toward the upper tail of the distribution.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

3.1.4. Distribution of model residuals

Figure 1c shows the residual distribution for the best-performing explanatory model. The residuals are centered close to zero and are approximately symmetric, with a mean of −0.004 and a standard deviation of 0.053. This indicates that the model has little average bias across the full evaluation set.

However, the distribution also has elongated tails. These larger residuals occur under a smaller subset of conditions and are consistent with the increased spread observed at higher AOD values in the observed–predicted comparison. For this reason, average error statistics alone are not sufficient to describe model performance across the full dust range.

3.1.5. Conditional bias across the AOD range

The mean residual provides a useful summary, but it does not show how error changes across the AOD distribution. Figure 1d therefore presents conditional bias as a function of predicted AOD. In the low-to-moderate range, conditional bias remains close to zero, indicating relatively stable calibration for the most common conditions.

At higher predicted AOD values, conditional bias becomes increasingly negative. Because residuals are defined as observed minus predicted in this diagnostic, negative values indicate underestimation in the upper part of the distribution. The model therefore reconstructs typical dust conditions more accurately than severe dust episodes. This range-dependent behavior motivates the separate analysis of extreme-event performance in the following sections.

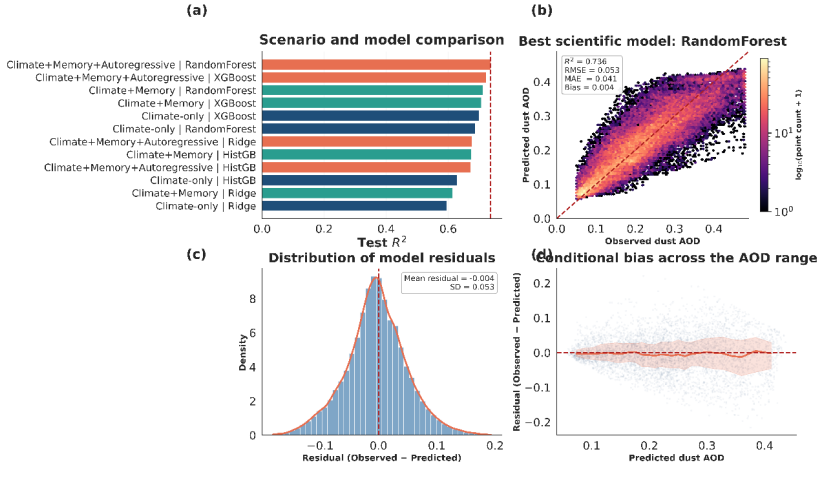


Figure 1. Overall evaluation of model performance in reconstructing dust AOD. (a) Test-set R^2across algorithms and explanatory configurations. (b) Observed versus predicted AOD for the best-performing explanatory model, Random Forest under the Climate+Surface+Memory configuration. The dashed line marks the 1:1 reference. (c) Distribution of residuals, defined as

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

observed minus predicted. (d) Conditional bias as a function of predicted AOD, showing how model error changes across the response range.

3.2. Added Value of Land-Surface Information

3.2.1. Improvement in skill after adding surface variables

Adding land-surface variables to the climate-only predictor set improved dust AOD reconstruction in a consistent way. At the domain scale, the Random Forest model improved from $R^2 = 0.686$ in the Climate-only configuration to $R^2 = 0.713$ in the Climate+Surface configuration. Over the same transition, $RMSE$ decreased from 0.058 to 0.056 and MAE decreased from 0.044 to 0.042. These changes correspond to an absolute gain of 0.027 in R^2, an absolute reduction of 0.002 in $RMSE$, and an absolute reduction of 0.002 in MAE. Bias changed only slightly, from 0.000 to -0.002.

These results indicate that instantaneous land-surface information adds explanatory power beyond atmospheric predictors alone. The gain is physically plausible because climate-only predictors do not directly represent surface resistance, vegetation shielding, or moisture-related changes in sediment availability. Surface variables provide complementary information on the state of erosion thresholds and the spatial structure of erodibility. In the benchmark logic of this study, this step defines the added value of the DMP-2 tier relative to DMP-1.

3.2.2. Spatial pattern of error reduction

The gain associated with surface variables is not spatially uniform. Figure 2 shows the difference in RMSE between the Climate-only and Climate+Surface configurations. Positive values indicate reduced error after the inclusion of surface predictors. The largest improvements are concentrated in transition areas and in regions with stronger vegetation influence. Smaller gains are found in more homogeneous desert interiors.

This spatial pattern suggests that surface information is most useful where soil moisture and vegetation vary more strongly over time and space. In such regions, climate-only predictors are less able to capture changes in effective dust-source conditions. By contrast, bare desert interiors tend to show more stable surface states. In those areas, the additional information carried by surface variables is smaller. From a benchmark perspective, these results show that DMP-2 is especially informative in regions where ecological transitions are active and surface-state dynamics are strong.

3.2.3. Regime-specific changes in error and bias

The regime-based results confirm that the impact of surface variables differs across benchmark environments. Relative RMSE reduction was largest in the vegetated zone, where error decreased by 2.89 %. The corresponding reductions were 1.67 % in the transition zone and

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

0.80 % in the core desert. Bias behavior was less uniform. The vegetated zone showed a 2.56 % reduction in bias, whereas the transition zone and core desert showed small bias degradations of −0.70 % and −1.57 %, respectively. These results suggest that surface variables mainly reduce random error across all regimes, but their effect on systematic bias is more region-dependent. In vegetated and semi-vegetated areas, surface predictors improve both overall fit and consistency. In very dry regions, they still reduce RMSE, but they may also introduce small residual bias. This behavior may reflect measurement limitations over bright desert surfaces or unresolved microtopographic effects.

Table 3. Relative changes in RMSE and Bias after adding instantaneous surface variables to the Climate-only configuration. Positive values indicate error reduction. Negative values indicate performance degradation.

Surface regime	RMSE change (%)	Bias change (%)
Vegetated zone	2.89	2.56
Transition zone	1.67	−0.70
Core desert	0.80	−1.57

3.2.4. Role of soil moisture and vegetation indicators

The spatial and regime-dependent gains are consistent with the physical role of soil moisture and vegetation in dust emission. Soil moisture increases particle cohesion and raises the effective erosion threshold. Vegetation increases roughness and sheltering, and it can also reduce sediment availability at the surface. These effects act directly on the susceptibility of the land surface to wind erosion.

The larger gains observed in vegetated and transition regions indicate that these controls are especially relevant where surface conditions are seasonally variable. They also help explain why nonlinear tree-based models performed better once surface predictors were included. The effect of wind depends on moisture and vegetation state. A model that can represent interactions

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

and threshold behavior is therefore better suited to this part of the problem.

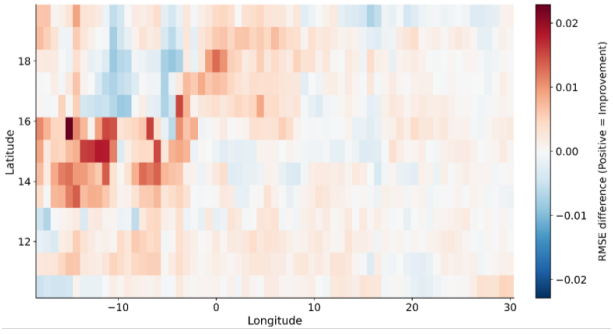
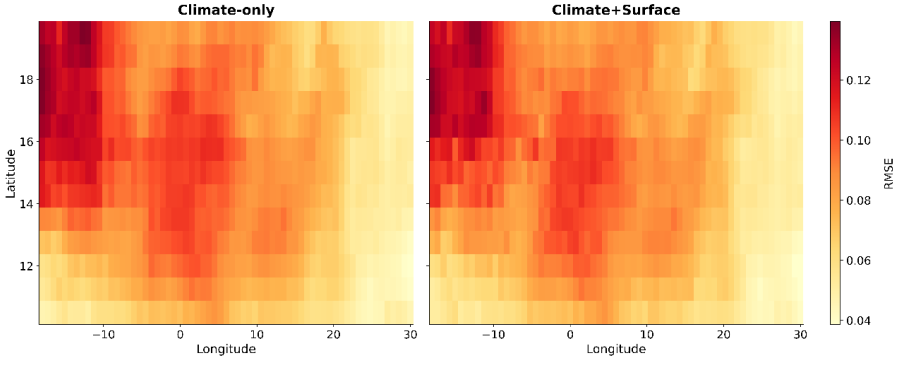


Figure 2. Difference in RMSE between the Climate-only and Climate+Surface configurations. Positive values indicate reduced error after the inclusion of instantaneous surface variables. The strongest gains occur in transition zones and in regions with stronger vegetation influence.

3.2.5. Benchmark interpretation across surface regimes

Taken together, the results show that the added value of surface variables is benchmark-region dependent. Vegetated and semi-vegetated zones show the strongest gains. Transition zones show intermediate but still meaningful improvement. Core desert regions show the smallest gain, which is consistent with their more stable surface conditions. This hierarchy is visible both in the regime statistics and in the spatial RMSE maps shown in Figure 3.

This pattern has a clear implication for the benchmark structure proposed in this study. Surface-state information should not be treated as an optional refinement that only improves local detail. It changes model skill in a systematic and regionally interpretable way. These findings justify the inclusion of an explicit surface-state experiment tier in a future DesertMIP protocol, particularly for transition and semi-vegetated regimes where climate-only predictors are least adequate.



https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 3. Spatial distribution of RMSE for the Climate-only configuration (a) and the Climate+Surface configuration (b). Error reduction is concentrated in benchmark regions with stronger land-surface variability, especially semi-arid and transition environments.

3.3. Role of Surface Memory in Improving Dust Reconstruction

3.3.1. Improvement in skill after adding memory components

Adding surface-memory terms produced a further improvement in dust AOD reconstruction after instantaneous surface variables had already been included. The increase in skill was modest, but it was consistent with the idea that dust activity is influenced not only by current land-surface conditions, but also by surface states inherited from previous months. Antecedent soil moisture and vegetation conditions may reflect accumulated drying, delayed vegetation response, or persistent exposure of erodible surfaces.

The improvement should not be overstated. The differences among the memory configurations are small, and they do not imply that a single lag window controls dust variability everywhere. Nevertheless, the direction of the change supports the inclusion of a separate memory-aware tier in the proposed benchmark hierarchy.

3.3.2. Sensitivity to lag length

Sensitivity analysis was carried out for lag windows of 0, 1, 3, and 6 months. All lagged-memory cases performed slightly better than the zero-lag case in terms of mean R^2. The zero-lag configuration gave mean $R^2 = 0.671$and mean $RMSE = 0.0576$. The 1-month and 3-month lags both yielded mean $R^2 = 0.678$and mean $RMSE = 0.057$. The 6-month lag gave the highest mean $R^2 = 0.679$and the lowest standard deviation of $R^2 = 0.0699$. It also gave the lowest standard deviation of $RMSE = 0.0043$ (Figure 4).

These results indicate that lagged predictors improve both mean skill and stability relative to the zero-lag baseline. At the same time, the differences among the lagged cases remain small. Importantly, the six-month lag gives the best predictive cross-validation performance in the present setup, whereas Sect. 3.7 identifies the strongest lagged associations near three months. This pattern supports the interpretation that antecedent surface conditions contain useful information, but it does not justify a strong claim that one lag window is uniquely optimal under all settings (Table 4).

Table 4. Cross-validation performance for different surface-memory lag lengths.

Lag (months)	Mean R²	SD(R²)	Mean RMSE	SD(RMSE)
0	0.671	0.0766	0.0576	0.0047

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Lag (months)	Mean R^2	SD(R^2)	Mean RMSE	SD(RMSE)
1	0.678	0.0786	0.0570	0.0051
3	0.678	0.0734	0.0570	0.0046
6	0.679	0.0699	0.0570	0.0043

3.3.3. Variation of model performance across lag windows

The lag-sensitivity results suggest that memory terms may help stabilize model performance. When lagged surface variables were included, mean skill increased slightly and the spread of performance across cross-validation folds generally declined, especially for the 3-month and 6-month windows. This behavior is consistent with the role of memory variables as temporal summaries of recent surface conditions, rather than as direct replacements for current-month predictors.

At the same time, the effect remains moderate. The results do not show a sharp separation among the tested lag windows. Instead, they indicate that antecedent surface information is useful within a short-to-seasonal time range. Together with the lag-association analysis in Sect. 3.7, these findings justify treating surface memory as a distinct diagnostic tier, while avoiding a strong claim that one lag length is universally optimal.

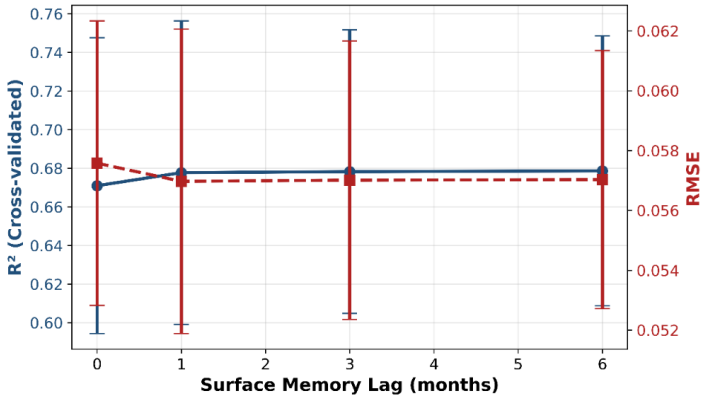


Figure 4. Cross-validation performance as a function of the lag applied to land-surface variables. Lagged-memory formulations show slightly higher mean R^2 and lower performance spread than the zero-lag baseline. The 6-month lag gives the best mean predictive performance in the present setup, although differences among lag windows remain modest. This should be distinguished from the lag-association analysis in Sect. 3.7, which peaks near three months.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

3.3.4. Spatial pattern of memory contribution

The contribution of memory terms is not spatially uniform. Feature-importance analysis based on SHAP indicates that memory effects are strongest in dry and semi-dry parts of the study region. These areas are more likely to retain multi-month signatures of drying, vegetation decline, and persistent surface exposure. By contrast, regions with denser vegetation and wetter surface conditions show weaker memory contributions. This spatial contrast is consistent with the expectation that antecedent surface state matters most where erosion sensitivity evolves gradually through time rather than responding only to the current month.

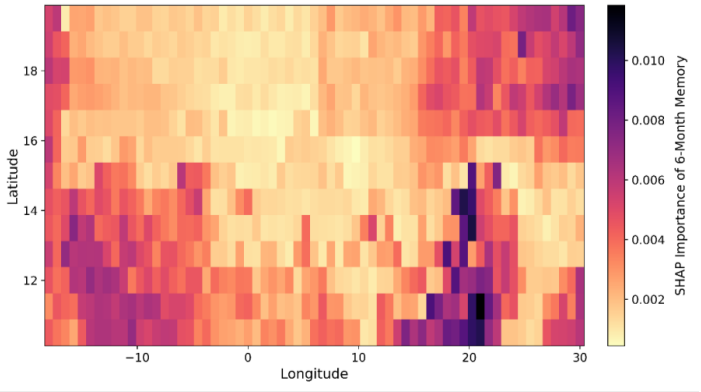


Figure 5. Spatial importance of land-surface memory based on SHAP values for the 6-month lag formulation. Higher values indicate regions where antecedent surface state contributes more strongly to the reconstruction of dust AOD. Memory effects are strongest in dry and semi-dry environments and weaker in more densely vegetated areas.

3.4. Model Explainability and Variable Importance

3.4.1. Global importance structure across model configurations

The explainability analysis shows that dust AOD variability is controlled by a combination of atmospheric forcing, instantaneous land-surface state, and antecedent surface memory. Figure 6 summarizes the importance structure across the three explanatory configurations. In the Climate-only configuration, atmospheric variables dominate by construction. Among these predictors, air temperature shows the highest importance, followed by wind-related terms and the spring seasonal indicator. This pattern suggests that the large-scale thermodynamic and dynamical state of the atmosphere exerts a strong influence on the background conditions that favor dust activity.

The inclusion of land-surface variables in the Climate+Surface configuration changes the structure of predictor importance. In particular, vegetation-related predictors such as leaf area index (LAI) become prominent. This shift indicates that part of the explained variability is

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

transferred from purely atmospheric drivers to variables that represent surface resistance and vegetation control. After memory terms are added, the importance structure expands further. In the full configuration, both instantaneous surface variables and memory-based predictors contribute to the explanation of dust variability. These results indicate that atmospheric forcing alone does not capture the full information content of the reconstruction problem.

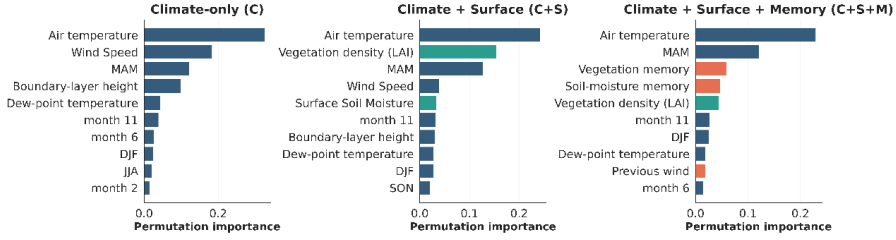


Figure 6. Permutation-based variable importance across the three explanatory configurations. Atmospheric variables dominate in the Climate-only configuration, while land-surface and memory variables become more important in the Climate+Surface and Climate+Surface+Memory configurations.

3.4.2. Shifts in variable-group contribution

The transition from Climate-only to Climate+Surface and then to Climate+Surface+Memory is associated with a redistribution of importance across predictor groups. In the Climate-only case, all explanatory power is assigned to atmospheric variables. Once surface variables are introduced, part of that importance shifts toward predictors that describe vegetation and moisture state. In the full configuration, memory variables account for an additional share of the explanatory structure. This result suggests that antecedent land-surface conditions contain information that is not fully captured by climate forcing or instantaneous surface state alone.

The grouped contribution estimates from the importance analysis are reported here only as relative importance shares, not as a second set of model-performance metrics. In the Climate+Surface configuration, climate variables account for about 74.45 % of the total importance, whereas surface variables account for about 25.54 %. In the Climate+Surface+Memory configuration, the corresponding shares are about 71.14 % for climate variables, 9.15 % for instantaneous surface variables, and 19.69 % for memory variables. These percentages indicate how the explanatory structure of the model changes as additional information is introduced. They should therefore be interpreted as diagnostic indicators of model dependence on different predictor groups, rather than as causal attribution or as substitutes for the performance metrics reported in Table 2.

3.4.3. Response-shape diagnostics for key predictors

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Partial-dependence plots provide additional insight into how the reference model responds to key predictors. Figure 7 shows that several predictors are associated with nonlinear and threshold-like behavior. Wind speed is associated with a rapid rise in predicted AOD at low to moderate values. The response then tends to flatten near about 5–6 m s^{-1}. This pattern is consistent with a transition from a strongly wind-limited regime to a more supply-limited regime.

Surface soil moisture shows a negative and nonlinear relationship with predicted AOD. The strongest decline occurs at very low moisture levels, roughly within the 0–0.05 range. At higher moisture values, the response becomes flatter. This pattern is physically plausible because small increases in moisture at the dry end can strongly increase particle cohesion, while further increases have less incremental effect on already stabilized surfaces.

Vegetation and memory predictors show similar response-shape behavior. Higher LAI values are associated with lower predicted AOD. The three-month soil-moisture memory term also shows a decreasing response. This result suggests that wetter antecedent conditions reduce the likelihood of dust activity during the current month. These plots should be interpreted as standardized response summaries of the fitted model rather than as direct causal evidence.

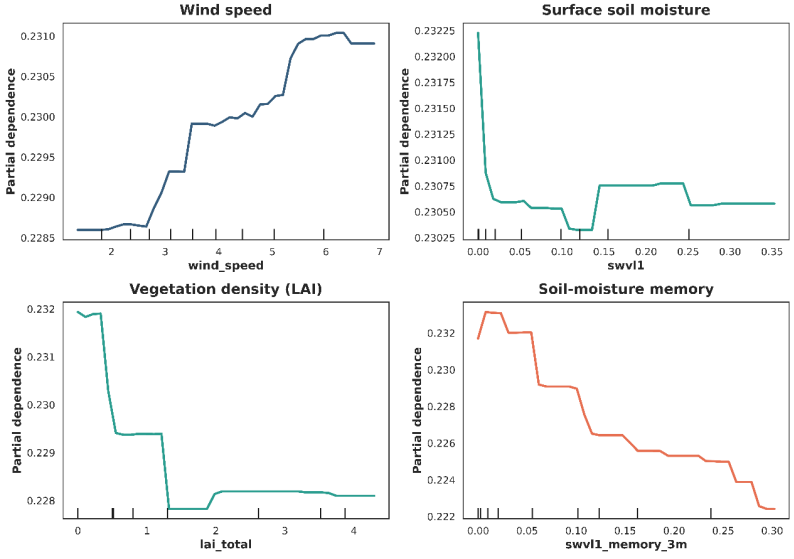


Figure 7. Partial-dependence plots for selected predictors, including wind speed, soil moisture, LAI, and soil-moisture memory. The plots show nonlinear response shapes and threshold-like behavior in the fitted model.

3.4.4. Interaction diagnostics in the erosion response

The interaction between wind speed and soil moisture provides a useful example of coupled

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

control in the fitted model. Figure S1 shows that high predicted AOD values occur mainly under the joint condition of strong winds and very low soil moisture. As soil moisture increases, predicted AOD declines even when wind remains relatively strong. This pattern is consistent with the idea that atmospheric forcing alone is not sufficient to explain erosion response. The land surface modulates whether strong wind can translate into high dust activity.

This interaction should be interpreted as a model-based diagnostic of coupled sensitivity rather than as a direct process proof. Even so, it provides useful interpretive support for the climate–surface–memory framework used in this study. Within a prototype DesertMIP framework, variable-group attribution and response-shape diagnostics are proposed as complementary interpretive products to accompany standard error metrics.

3.5. Regime-based Analysis of Error and Bias

Regime-based evaluation is treated here as a core DesertMIP diagnostic because dust-model behaviour is expected to differ systematically between desert cores, ecological transition zones, and more vegetated margins. In this section, mean bias is defined consistently as predicted minus observed. Under this convention, positive bias indicates overestimation and negative bias indicates underestimation.

3.5.1. Error distribution across core desert, transition, and vegetated regimes

Evaluation of the final Climate+Surface+Memory configuration shows that prediction error varies across benchmark surface regimes. The highest RMSE occurs in the transition zone (0.078), followed by the core desert (0.076), whereas the vegetated zone shows the lowest error (0.067). This pattern indicates that transition environments are the most difficult regime for dust AOD reconstruction. The elevated uncertainty in transition zones is consistent with their stronger environmental heterogeneity. These regions combine sharper gradients in vegetation, moisture, and surface resistance than the more homogeneous desert core. By contrast, the vegetated zone shows the lowest error level, which suggests that model behaviour is more stable where land-surface conditions are comparatively less dust-prone and more structured. The core desert occupies an intermediate position, with error levels lower than in the transition zone but higher than in the vegetated regime.

3.5.2. Spatial concentration of bias

The regional bias map for the final configuration is shown in Figure 8. Because bias is defined here as predicted minus observed, positive values indicate overestimation. Red areas therefore mark regions where the model tends to predict dust AOD values that are too high. The map shows that positive bias is spatially concentrated in transition environments and along desert margins. Mean bias values are positive in all three regimes, but they differ only modestly: the transition zone shows the largest mean bias (0.015), followed by the core desert (0.014) and the

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

vegetated zone (0.013). These results indicate that transition environments are not only the most uncertain in terms of RMSE, but also the most sensitive to systematic error. At the same time, the relatively small differences among the three regimes suggest that regime contrasts are more pronounced for error magnitude than for mean bias alone. This behaviour is consistent with the stronger ecological and climatic gradients that characterize transition regions.

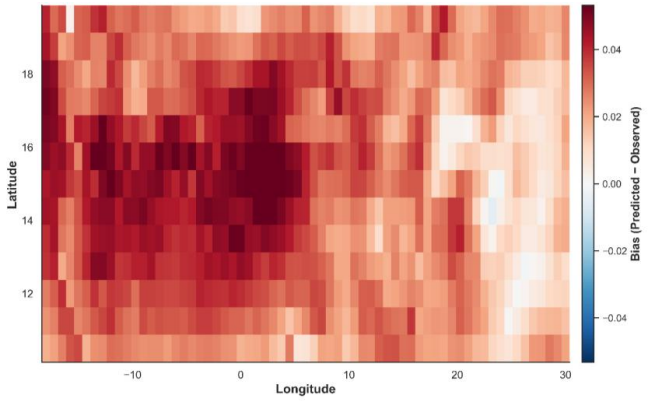


Figure 8. Spatial distribution of model bias for the Climate+Surface+Memory configuration. Bias is defined as predicted minus observed. Positive values (red) indicate overestimation, and negative values (blue) indicate underestimation. Overestimation is concentrated mainly in transition environments and along desert margins.

3.5.3. Relationship between bias and the Desert Proxy Index

Figure 9a examines the relationship between model bias and the Desert Proxy Index. The overall association is weak. The Pearson correlation coefficient is $r = -0.02$, and the Spearman coefficient is $\rho = -0.08$. These values indicate that there is no strong monotonic relationship between mean bias and increasing desert-like conditions across the full domain.

Even so, the hexbin distribution and the median trend line show greater dispersion of positive bias values within the Desert Proxy Index range of about 0.6 to 0.9. This pattern suggests that highly arid environments can still be associated with less stable model behaviour, even though the relationship is not strong in a domain-wide linear sense. The result therefore points to localized sensitivity rather than to a simple large-scale bias gradient.

3.5.4. Relationship between RMSE and transition intensity

Figure 9b shows the relationship between RMSE and transition intensity. The association is again weak, with a Pearson correlation coefficient of $r = -0.06$. Mean RMSE values across transition-intensity classes are also very similar. The mean RMSE is 0.081 for low intensity, 0.078 for moderate intensity, and 0.079 for high intensity. These results indicate that transition intensity alone does not explain error magnitude in a simple linear way. At the same time, the

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

regime-based comparison in Sect. 3.5.1 still shows that transition zones have the highest overall RMSE. The implication is that transition-related uncertainty is better expressed at the regime level than through a simple pointwise correlation with transition intensity.

3.5.5. Changes in spatial error patterns across model configurations

The comparison across model configurations shows that adding surface and memory predictors improves the spatial structure of model error, although the improvement is not uniform across regimes. In the final Climate+Surface+Memory configuration, RMSE is 0.076 in the core desert, 0.078 in the transition zone, and 0.067 in the vegetated zone (Table 7). The transition zone therefore remains the most difficult environment to reconstruct, while the vegetated zone has the lowest residual error.

This ordering is consistent with the idea that desert margins and transition environments are more difficult to model because they combine sharp gradients in vegetation, soil moisture, and surface resistance. Figure S2 further shows that positive-bias clusters are stronger in the Climate-only configuration and become weaker after surface and memory information is included. Thus, the full configuration improves not only the average skill, but also the regional coherence of the error field.

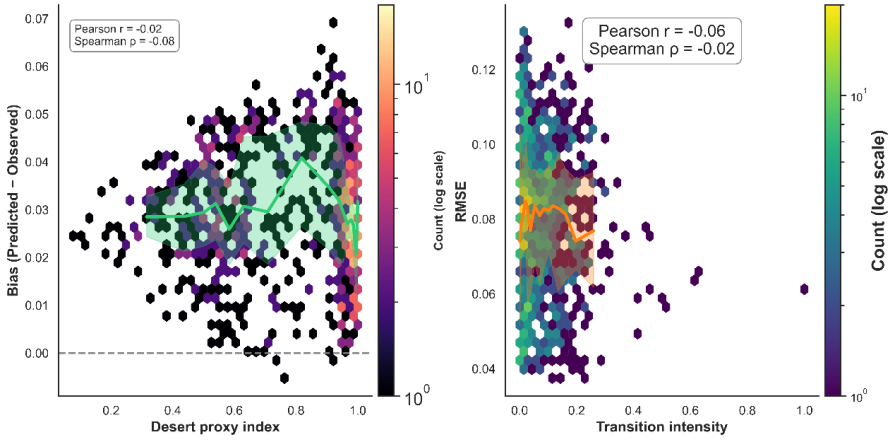


Figure 9. Relationships between model error metrics and surface indices. (a) Relationship between model bias, defined as predicted minus observed, and the Desert Proxy Index. The hexbin shading shows point density, and the thick line shows median bias across index intervals. (b) Relationship between RMSE and transition intensity. Mean RMSE varies only weakly across transition-intensity classes.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

3.6. Model Performance in the Reconstruction of Extreme Dust Events

3.6.1. Evaluation of extreme-event detection

Model performance under extreme conditions was evaluated using the upper 5 % of the AOD distribution. This subset corresponds to 2,231 samples in the test dataset and represents severe dust events. Detection metrics for these events are summarized in Table 5. The results show a clear reduction in skill relative to the overall reconstruction problem. In particular, the probability of detection (POD) and the critical success index (CSI) are lower than would be expected from the aggregate skill scores alone. This pattern indicates that severe dust outbreaks are more difficult to reconstruct than moderate or typical conditions.

This result is physically and statistically plausible. Extreme events are rare. They also tend to occupy the upper tail of the response distribution, where regression models often show stronger shrinkage toward central values. As a result, acceptable mean-state performance does not guarantee accurate representation of severe dust outbreaks.

3.6.2. Comparison of model configurations for extreme events

A comparison of the three explanatory configurations shows that the inclusion of surface and memory predictors does not improve extreme-event detection. Instead, detection metrics decline slightly from the Climate-only configuration to Climate+Surface and then to Climate+Surface+Memory. POD decreases from 0.437 in the Climate-only case to 0.431 in Climate+Surface and to 0.423 in Climate+Surface+Memory. CSI decreases from 0.276 to 0.272 and then to 0.269. FAR remains high in all cases, with values between 0.572 and 0.576.

These results indicate that the added value of surface and memory predictors is concentrated in the reconstruction of the overall distribution and mean-state structure rather than in the detection of the most severe events. From a DesertMIP perspective, this result is valuable because it identifies tail performance as a separate diagnostic target rather than assuming that improvements in overall skill automatically transfer to severe dust outbreaks.

Table 5. Detection metrics for extreme dust events across the three explanatory configurations. Extreme events are defined as observations above the 95th percentile of the AOD distribution.

Scenario	POD	CSI	FAR
C	0.437	0.276	0.572
C+S	0.431	0.272	0.576
C+S+M	0.423	0.269	0.574

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

3.6.3. Error amplification during extreme events

The error structure changes markedly under extreme conditions. For the full Climate+Surface+Memory configuration, the extreme-event RMSE is 0.170 and the extreme-event MAE is 0.141, with an extreme bias of −0.060. Under the bias convention used here, this negative value indicates systematic underestimation of extreme dust intensity. Similar behavior is also found in the Climate-only and Climate+Surface configurations, which yield extreme bias values of −0.058 and −0.059, respectively.

Taken together, these results show that the upper tail of the AOD distribution remains difficult to reconstruct across all configurations. While surface and memory information improve the physical realism of the framework and support better process representation, they do not eliminate the tendency toward underestimation in the most severe events. In other words, the main remaining challenge lies less in identifying where extreme events occur than in reproducing their full magnitude.

Table 6. Quantitative error statistics for extreme dust events across the three explanatory configurations

Scenario	n extreme	Extreme RMSE	Extreme MAE	Extreme Bias
C	2231	0.161	0.133	−0.058
C+S	2231	0.163	0.134	−0.059
C+S+M	2231	0.170	0.141	−0.060

3.6.4. Implications for benchmark design and future model development

Figure 10 shows that the full model reproduces the broad geographic pattern of extreme dust occurrence, but it still underestimates event intensity. This distinction is important because spatial placement and event magnitude are not the same measure of skill. A model may identify the regions where severe dust is likely to occur while still failing to reproduce the strength of the most intense events.

For this reason, severe-event evaluation should be treated separately from mean-state reconstruction. The present results do not show that surface-memory terms improve extreme-event detection. Future model development may therefore require methods that are more sensitive to rare events, such as tail-aware loss functions, resampling strategies, or specialized extreme-value diagnostics.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

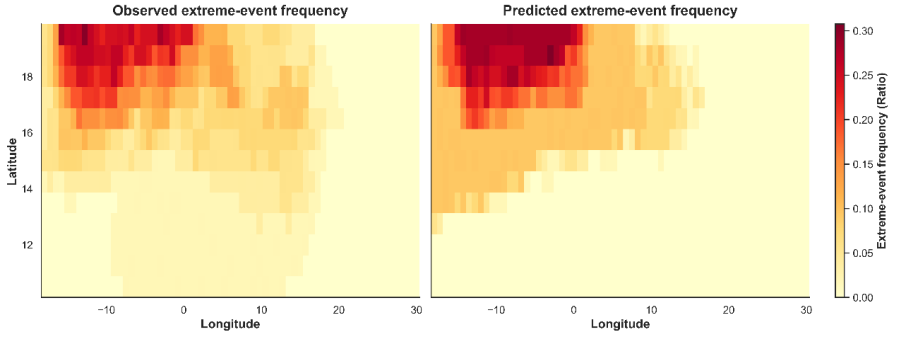


Figure 10. Spatial distribution of extreme dust AOD events for the observed data and for the Climate+Surface+Memory reconstruction. The model reproduces the broad spatial structure of extreme-event occurrence but underestimates event intensity, especially in the most severe cases.

3.7. Surface Memory Analysis and Lagged Relationships

3.7.1. Lag-dependent associations between surface variables and AOD

To further examine the temporal structure of surface influence, Pearson correlations were calculated between dust AOD and two key land-surface variables, soil moisture (SM) and leaf area index (LAI), across lags from 0 to 6 months. The results are shown in Figure 11. Correlation strength varies systematically with lag. Contemporaneous correlations are relatively weak, with values of about -0.09 for soil moisture and -0.18 for LAI at lag 0. Stronger negative correlations emerge at longer lags, and the largest magnitudes occur near a lag of about 3 months. Peak values reach about -0.75 for soil moisture and -0.82 for LAI. Correlation strength then weakens after the fourth month.

These results indicate that lagged surface associations are stronger than contemporaneous associations in the present dataset. Importantly, these association peaks near three months should be distinguished from the predictive-skill comparison in Sect. 3.3, where the six-month lag yielded the highest mean cross-validation performance, albeit by a small margin. However, these patterns should be interpreted with caution. They provide evidence for memory-sensitive benchmarking, not direct proof of causal memory mechanisms. These lag diagnostics are proposed as benchmark descriptors within the prototype DesertMIP framework, while recognizing that part of the signal may reflect seasonality and temporal autocorrelation.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

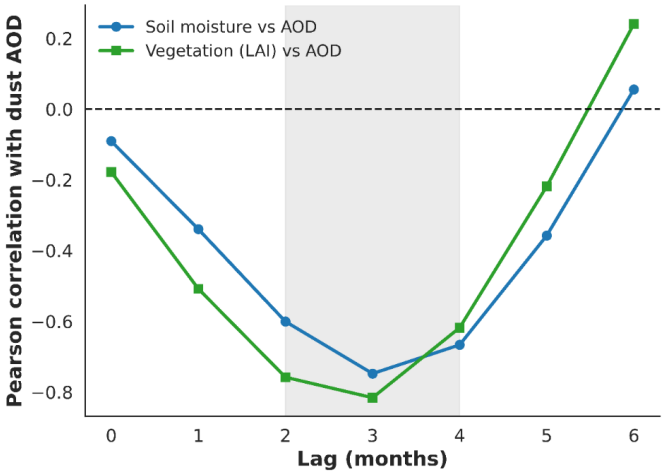


Figure 11. Pearson correlation between dust AOD and key land-surface variables, soil moisture (SM) and leaf area index (LAI), across temporal lags from 0 to 6 months. The strongest negative associations occur within the two- to four-month window, with a peak near three months. This lag of maximum association is distinct from the predictive-skill analysis in Sect. 3.3, in which the six-month lag produced the highest mean cross-validation score. These lag patterns are presented as benchmark descriptors of memory-sensitive behavior rather than as stand-alone proof of causal memory mechanisms.

3.7.2. Comparison of vegetation and soil-moisture behavior

The lag structure differs between LAI and soil moisture. LAI shows slightly stronger negative associations with dust AOD than soil moisture across the main lag window. Its maximum correlation reaches about -0.82, whereas the maximum for soil moisture is about -0.75. The decline in correlation strength at longer lags is also more gradual for LAI than for soil moisture.

This contrast is consistent with the idea that vegetation-related controls may persist longer than moisture-related controls in the present dataset. Vegetation affects roughness, sheltering, and surface stabilization. Soil moisture mainly affects particle cohesion. These differences may help explain why LAI retains a stronger lagged signal over a wider temporal window. Even so, the present results should be read as comparative lag associations rather than as direct evidence that one process is mechanistically dominant.

3.7.3. Temporal interpretation of lagged effects

The lag structure can be summarized conceptually by writing dust AOD as a function of both present and antecedent surface states:

$$AOD_t = f(S_t, S_{t-1}, S_{t-2}, \ldots) \quad (32)$$

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

where S denotes land-surface state. In the present dataset, the strongest associations occur within the two- to four-month window prior to the target month, with the most pronounced signal near three months. This association-based result is compatible with, but not identical to, the predictive experiment reported in Section 3.3, in which the six-month lag provided the highest mean cross-validation skill. The key point here is not that a unique causal lag has been identified. Rather, the results indicate that benchmark diagnostics based on antecedent surface state are informative and should be retained in a climate–surface–memory evaluation framework. For this reason, lag-sensitive association analysis is treated here as part of the prototype DesertMIP benchmark logic.

3.7.4. Surface-memory signatures under extreme dust conditions

The lag analysis also provides useful context for severe dust conditions. Extreme events were defined as the upper 5 % of the AOD distribution. Surface conditions associated with these events are summarized in Figure S3. Mean soil moisture during extreme events is about 0.026, compared with about 0.094 under non-extreme conditions. Vegetation indicators are also lower during extreme events, and the corresponding memory terms show similar reductions in the months preceding those events.

These differences indicate that severe dust conditions are associated with drier and less vegetated antecedent surface states. However, this result should not be interpreted as proof that surface memory alone improves extreme-event reconstruction. Sect. 3.6 showed that the addition of surface and memory predictors did not improve extreme-event detection metrics. The present subsection therefore identifies lagged surface contrasts associated with severe events, but it does not overturn the earlier conclusion that tail reconstruction remains limited.

3.7.5. Differences in lag response across benchmark regimes

Lag-dependent associations also vary across benchmark regimes, as shown in Figure 12. In the core-desert regime, the strongest correlations occur near a lag of three months, with values of about -0.49 for soil moisture and -0.70 for LAI. Stronger lagged associations appear in transition zones, where the corresponding values reach about -0.63 for soil moisture and -0.75 for LAI. Vegetated regions show a particularly strong LAI signal during the second and third months, with correlations exceeding about -0.71 in some cases.

These results indicate that lag-sensitive benchmark behaviour is regime dependent. Transition zones and more variable vegetated margins show stronger lag structure than the more homogeneous desert core. This pattern is consistent with the broader finding that surface-state and memory information are most informative where the land surface changes more strongly through time. As in the domain-wide analysis, these regime-based correlations identify the lag of strongest association, not necessarily the lag of best predictive skill. In summary, dust variability in the present dataset is associated with both instantaneous and antecedent surface

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

conditions. These lag diagnostics therefore support the inclusion of memory-sensitive descriptors in future DesertMIP-style benchmark evaluations, while remaining distinct from claims of direct causal attribution.

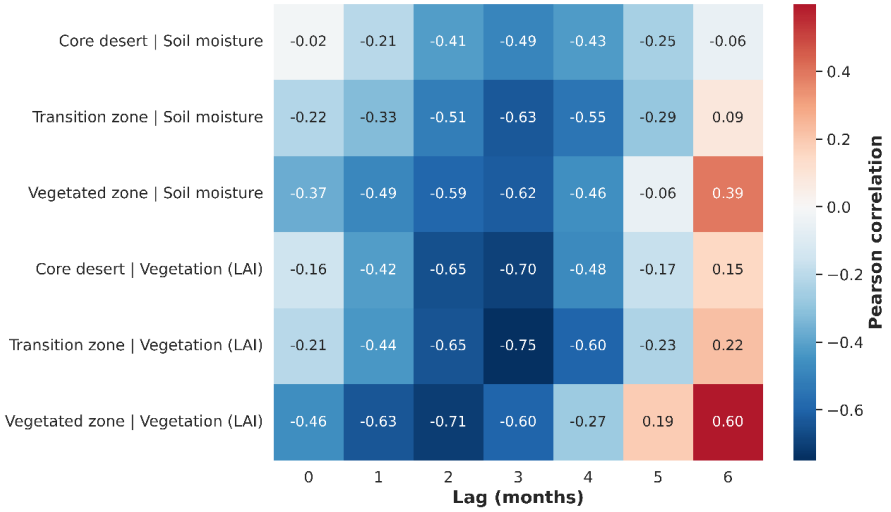


Figure 12. Lag-dependent associations between land-surface variables and dust AOD across benchmark regimes, including the core desert, transition zone, and vegetated region. The strongest lag structure appears in transition and more variable vegetated environments. The figure summarizes regime-dependent association peaks rather than predictive-skill optima. The figure is intended to show regime-dependent benchmarking signals within the climate–surface–memory framework.

3.8. Spatial Analysis of Dust Patterns and Model Error

Observed, reconstructed, and error maps are treated here as mandatory benchmark products because they reveal whether model gains are regionally coherent or confined to limited parts of the domain. In this section, spatial diagnostics are evaluated for the full Climate+Surface+Memory configuration. Consistent with Sect. 3.5, bias is defined as predicted minus observed. Positive values therefore indicate overestimation, and negative values indicate underestimation.

3.8.1. Spatial pattern of observed AOD

The multi-year mean map of observed dust AOD is shown in Figure 13a. The highest dust burden occurs in the northern and central parts of the study area. A broad core band between about 15° N and 20° N shows the largest values. Mean AOD reaches about 0.4 in the most active source regions. Southward from this band, AOD decreases as vegetation cover increases and surface conditions become less favorable for sustained dust emission.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

This spatial structure is consistent with the benchmark regime framework used throughout the study. It reflects the expected sequence from desert core to transition zone and then to more vegetated areas. For this reason, the observed map provides the reference spatial benchmark against which reconstructed patterns and model errors are evaluated.

3.8.2. Spatial pattern of reconstructed AOD

The reconstructed AOD field from the full Climate+Surface+Memory configuration is shown in Figure 13b. The model reproduces the main regional structure of dust activity. In particular, it captures the location of the main source band and the broad north–south gradient in mean AOD. This agreement indicates that the model is able to recover the dominant spatial organization of dust variability over North Africa.

At the same time, the comparison between observed and reconstructed means suggests that the model tends to produce slightly higher values over part of the domain. This tendency is modest in magnitude, but it is regionally widespread. The spatial error field provides a clearer view of where this bias is concentrated.

3.8.3. Spatial structure of model error

The spatial error map is shown in Figure 13c and is defined as predicted minus observed. Under this convention, positive values indicate overestimation and negative values indicate underestimation. Most grid-cell differences fall within about −0.03 to +0.03. Positive values are more widespread than negative values, which indicates that mild overestimation is the dominant error mode in the full model configuration.

The strongest positive errors are concentrated along desert margins and in parts of the transition zone. More localized negative errors occur in some central and eastern areas. This pattern suggests that the main limitation of the model is not a failure to reproduce the overall spatial structure of dust AOD. Instead, it is a tendency toward modest but spatially coherent bias in specific benchmark regions.

3.8.4. Regime-based summary of observed, reconstructed, and bias patterns

The available regime-based error metrics indicate that the vegetated zone has the lowest reconstruction error, with $MAE = 0.047$and $RMSE = 0.067$. The transition zone remains the most difficult regime, with the highest $RMSE = 0.078$. The desert core shows intermediate behavior, with $MAE \approx 0.053$. These results are consistent with the benchmark interpretation developed in Sect. 3.5. They suggest that model performance depends strongly on structural regime differences in land-surface state and heterogeneity (Table 7).

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Table 7. Regime-based spatial summary of observed, reconstructed, and bias fields for the full Climate+Surface+Memory configuration

Surface regime	Mean observed AOD	Mean predicted AOD	Mean bias (Predicted − Observed)	MAE	RMSE
Core desert	0.275	0.289	0.014	0.054	0.076
Transition zone	0.256	0.271	0.015	0.055	0.078
Vegetated zone	0.178	0.192	0.013	0.047	0.067

3.8.5. Relationship between spatial error and surface properties

The relationship between model error and surface characteristics is further examined in Figure S4. The spatial comparison suggests two broad patterns. First, larger absolute errors tend to occur in areas with stronger desert characteristics. Second, regions with stronger surface-memory signals tend to exhibit lower absolute errors. Together, these patterns indicate that model error is not randomly distributed, but is linked to spatial variations in aridity and antecedent land-surface memory.

These patterns are consistent with the broader benchmark interpretation of the study. Even in the full Climate+Surface+Memory configuration, the most difficult areas remain those where surface processes are highly threshold-dependent, spatially heterogeneous, or imperfectly constrained by available observations. From a DesertMIP perspective, such spatial diagnostics are useful because they identify where model gains are robust and where benchmark performance remains limited.

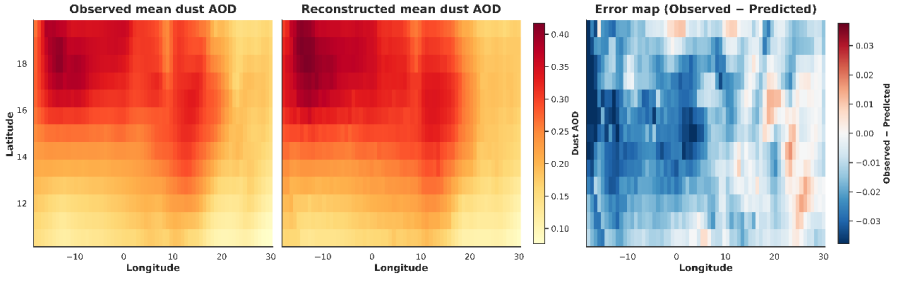


Figure 13. Mandatory benchmark maps for the full Climate+Surface+Memory configuration. (a) Multi-year mean observed dust AOD. (b) Multi-year mean reconstructed dust AOD. (c)

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Spatial error field, defined as predicted minus observed. Positive values indicate overestimation, and negative values indicate underestimation. Together, these maps show whether model performance is regionally coherent across the benchmark domain.

3.9. Uncertainty and Stability Analysis

In the context of a future DesertMIP protocol, sensitivity tests to variable blocks, temporal aggregation, threshold choice, and memory-window definition are recommended as minimum robustness checks for any benchmark implementation. The purpose of this section is therefore not to claim definitive robustness, but to evaluate whether the main findings remain qualitatively stable under several alternative analysis settings.

3.9.1. Sensitivity to variable selection

Feature-ablation experiments were used to test the dependence of model skill on major predictor blocks. The results shown in Figure 14 indicate three consistent patterns. First, removal of soil and vegetation variables reduces model performance. Second, removal of the memory block produces the largest decline in $R^2$and the largest increase in RMSE. Third, omission of surface-memory terms reduces model skill to a level close to that of the Climate-only configuration. These results suggest that the gain in the full configuration does not arise solely from redundant predictors. Instead, it reflects additional information associated with dynamic surface state and antecedent surface history.

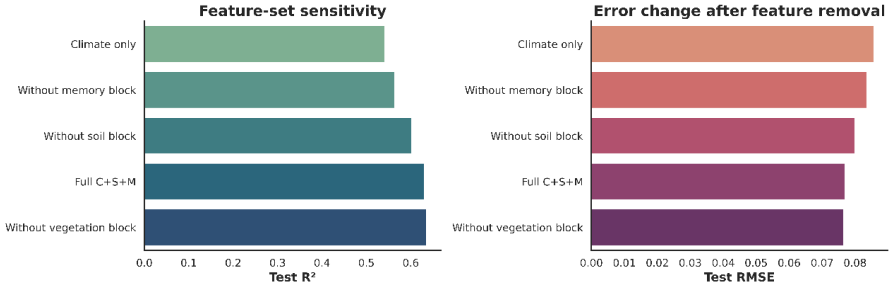


Figure 14. Sensitivity of model performance to variable-block removal. The left panel shows changes in R^2, and the right panel shows changes in RMSE after systematic removal of soil, vegetation, and memory predictor blocks. The figure highlights the contribution of surface-memory information to the full reconstruction framework.

3.9.2. Sensitivity to spatial and temporal scale

Sensitivity tests across temporal and spatial scales are shown in Figure 15. Temporal aggregation from monthly to seasonal resolution increases apparent skill, with R^2rising from about 0.55 to values above 0.80 and RMSE decreasing accordingly. This pattern is consistent with the reduction of short-term temporal noise under seasonal averaging. The relative

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

advantage of the Climate+Surface+Memory configuration remains most evident at the native spatial scale. This result suggests that surface and memory information are especially relevant for reconstructing finer-scale and shorter-term variability. At the same time, these comparisons should be interpreted as sensitivity analyses rather than as evidence of fully independent model generalization.

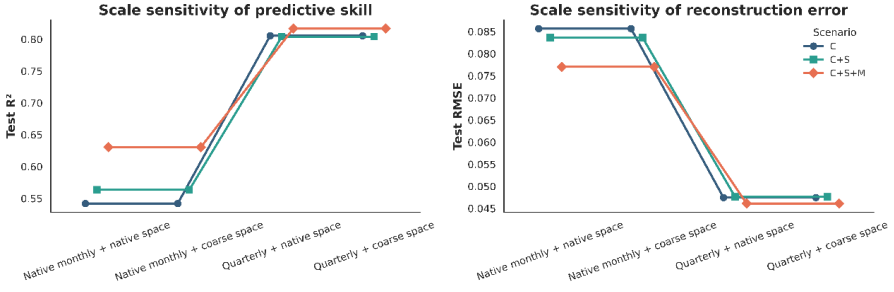


Figure 15. Sensitivity of model performance to spatial and temporal scale. The left panel shows changes in R^2, and the right panel shows changes in RMSE across alternative aggregation scales. Temporal aggregation increases apparent skill, while the advantage of the Climate+Surface+Memory configuration remains most visible at the native scale.

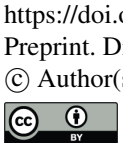
3.9.3. Stability across independent time periods

Model performance was also compared across two non-overlapping periods, referred to here as Early and Late. As shown in Figure 16, performance is higher in the Late period, with $R^2 \approx 0.75$, than in the Early period, with $R^2 \approx 0.44$. This difference may reflect changes in data quality, evolving surface conditions, or temporal variation in predictor–response relationships. Despite this shift in absolute performance, the ordering of the three explanatory configurations remains unchanged:

$$\text{Climate-only} < \text{Climate+Surface} < \text{Climate+Surface+Memory}$$

This consistency suggests that the relative benefit of surface and memory information is not confined to a single part of the record. However, the comparison is temporal rather than spatial, and it therefore does not remove the possibility that some optimism remains due to spatial autocorrelation.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.



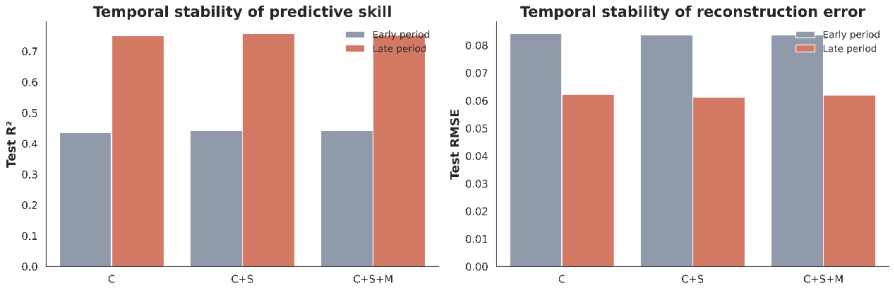


Figure 16. Stability of scenario ranking across two independent time periods. Performance is shown using $R^2$and RMSE for the Early and Late periods. Although absolute skill differs between the two periods, the ordering of the three explanatory configurations remains unchanged.

3.9.4. Inter-subregion variability

Figure 17 summarizes model behavior across the major surface regimes. The transition zone shows the highest RMSE, while the core desert follows closely behind. The vegetated region shows the lowest reconstruction error. The figure also indicates that the spread of absolute error is largest in the transition zone and remains similarly high in the core desert, whereas it is clearly smaller in the vegetated regime. This pattern is consistent with the regime-based results reported in Section 3.5 and Table 7, which identify transition environments as the most difficult benchmark regions and vegetated areas as the most stable ones. Together, these regime-level comparisons provide descriptive evidence of spatial heterogeneity in model performance. They should not, however, be interpreted as a substitute for stricter spatial validation such as leave-region-out testing.

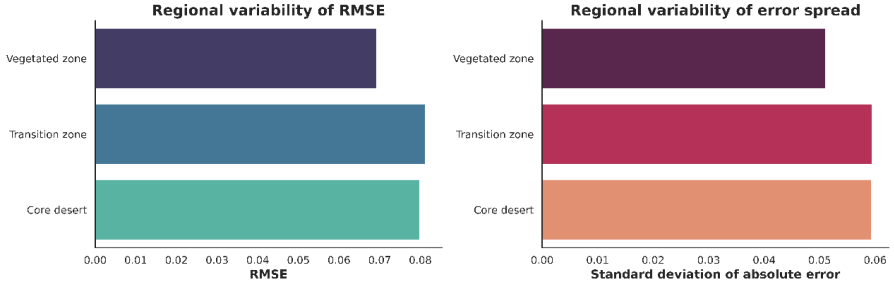


Figure 17. Regime-level summary of model error characteristics across the vegetated zone, transition zone, and core desert. The transition zone exhibits the highest RMSE, whereas the core desert shows a similarly large error spread. The vegetated zone has the lowest overall error levels. These contrasts highlight the spatial heterogeneity of model performance and confirm that transition environments remain the most challenging benchmark regime for dust-AOD

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

reconstruction.

3.9.5. Robustness across threshold and memory-window choices

Figure 18 examines two additional sensitivity dimensions: the choice of extreme-event threshold and the definition of the memory window. The left panel shows that stricter thresholds, from quantile 0.90 to 0.97, reduce event-detection scores. This pattern is expected because more severe tail events are rarer and harder to identify. The right panel shows that changing the memory window from two to four months produces only minor changes in model skill. Together, these results suggest that the main conclusions about the usefulness of surface and memory information are not highly sensitive to moderate changes in tuning choices.

Importantly, these robustness results should not be interpreted as evidence that the Climate+Surface+Memory configuration improves all aspects of performance under all thresholds. Section 3.6 showed that overall improvements in reconstruction skill do not automatically translate into better extreme-event detection, and that tail metrics such as POD and CSI can decline even when aggregate skill improves. The present sensitivity tests therefore support the stability of the general climate–surface–memory framework, but they do not overturn the earlier conclusion that severe-event reconstruction remains a separate challenge.

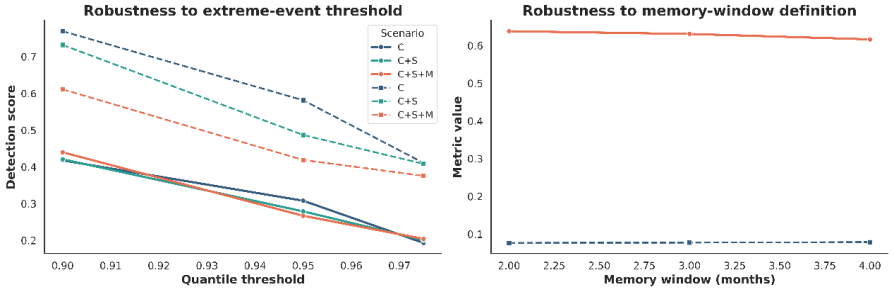


Figure 18. Robustness checks for the prototype benchmark framework. The left panel shows the sensitivity of extreme-event detection scores to the choice of quantile threshold. The right panel shows the sensitivity of overall model skill to the definition of the memory window from two to four months. The figure is intended as a protocol-level stability check rather than as evidence that all diagnostics improve uniformly under the full configuration.

3.9.6. Summary of robustness checks

Taken together, the sensitivity analyses support three main conclusions. First, the direction of the main findings remains broadly stable across variable-removal tests, scale changes, temporal partitions, and memory-window choices. Second, the addition of surface and memory information retains a consistent advantage in overall explanatory skill across these settings. Third, the present evidence should be interpreted as supportive rather than definitive because

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

the validation design is still primarily temporal and not fully spatially independent. In this sense, the robustness analysis provides a useful protocol-level check for the proposed benchmark framework, while also identifying the need for stricter spatial validation in future DesertMIP-style implementations.

4 Discussion

4.1. What is robustly supported by the present results

The present results support four main conclusions. First, climate variables alone are not sufficient to explain the observed variability of dust AOD over North Africa. The addition of dynamic land-surface variables improves reconstruction skill in a consistent way. This result agrees with earlier studies that identified wind as a necessary but not sufficient control on dust emission. Soil moisture, vegetation cover, surface roughness, and source characteristics also shape erosion thresholds and emission efficiency (Mousavi et al., 2023; Wu et al., 2016). The present study does not introduce entirely new controls. Its contribution is to evaluate them within a single hierarchical framework and to quantify their incremental value step by step.

Second, lagged surface predictors add a further, although modest, improvement beyond instantaneous surface variables. This result is consistent with the idea that dust variability is not fully captured by a purely synchronous representation of land-surface state. Instead, part of the signal appears to be associated with antecedent conditions such as progressive drying, vegetation decline, and multi-month changes in surface resistance. This interpretation is broadly consistent with studies from the Sahel and the Sahara, which show that rainfall, soil moisture, and vegetation can influence dust activity over both short and extended timescales (Bergametti et al., 2017). At the same time, the lag analysis in the present study should be interpreted cautiously. It supports the usefulness of memory-sensitive predictors and diagnostics, but it does not by itself isolate a unique causal mechanism.

Third, regime dependence is strong. Model error is not distributed uniformly across the domain. Transition zones show the highest uncertainty and the strongest sensitivity to surface and memory information. This pattern is physically plausible. Transitional environments combine sharper gradients in vegetation, moisture, and surface roughness than the more homogeneous desert core. Recent work has also emphasized the importance of spatial heterogeneity and resolution in dust modeling (Chappell et al., 2024). In that context, the present results support the broader view that fixed source masks and static boundary descriptions can be least adequate where ecological boundaries are dynamic (LeGrand et al., 2023).

Fourth, severe events remain difficult to reconstruct. The model reproduces large-scale spatial structure and mean behaviour more successfully than the upper tail of the AOD distribution. This limitation is consistent with earlier evaluations of operational and research dust models,

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

which often reproduce broad patterns more successfully than peak magnitude or event frequency (L. Zhang et al., 2022). In the present case, the addition of surface and memory predictors improves overall reconstruction skill, but it does not improve extreme-event detection metrics. This distinction is important and should be retained throughout the interpretation of the study.

4.2. What do these findings mean for DesertMIP?

The present findings support the need for a benchmark structure in which dust-model performance is evaluated not only against overall error metrics, but also with respect to surface-state sensitivity, regime dependence, antecedent-memory effects, and tail behaviour. In this sense, the current study provides a pilot diagnostic blueprint toward DesertMIP.

This implication arises directly from the structure of the results. The hierarchical comparison shows that the step from climate-only predictors to climate plus surface predictors is informative. The further step from instantaneous surface state to surface memory also remains informative, even though the quantitative gains are modest. Regime-based analyses show that these gains are not spatially uniform. Extreme-event analyses show that improvements in mean-state performance do not automatically transfer to severe outbreaks. Taken together, these results point to a benchmark logic that is more structured than a single global score.

From this perspective, DesertMIP should not be understood simply as another intercomparison label. It should be understood as a framework for organizing dust-model evaluation around the climate–surface–memory problem. That problem includes at least four linked dimensions: overall skill, regime-specific behaviour, lag-sensitive surface influence, and tail performance. The present study suggests that all four dimensions are necessary if dust-model behaviour is to be evaluated in a way that is both physically interpretable and diagnostically useful.

4.3. What should a DesertMIP-style benchmark contain?

A useful DesertMIP-style benchmark should include a small number of clearly defined elements. First, it should include hierarchical experiments. At minimum, these should distinguish climate-only predictors, climate plus instantaneous surface state, and climate plus surface plus memory. Such a structure allows the added value of each information block to be tested directly.

Second, it should use public and reproducible benchmark inputs. This is important for transferability and for independent repetition of the benchmark across models and groups. Third, it should include regime masks that separate desert cores, transition zones, and more vegetated margins. The present results show that these zones behave differently and should therefore be treated as mandatory reporting strata rather than optional regional summaries.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Fourth, the framework should include tail diagnostics in addition to standard regression metrics. The present study shows clearly that aggregate skill and extreme-event skill are not interchangeable. A model may improve in $R^2$and RMSE while still failing to improve POD, CSI, or extreme-event bias. Fifth, the framework should retain lag diagnostics because they provide a useful description of memory-sensitive behaviour, even if they do not by themselves establish causality. Finally, a DesertMIP-style benchmark should include reproducible scripts, experiment definitions, and standardized output products. Without these elements, the framework remains conceptual rather than operational.

4.4. What is not yet demonstrated

The present study does not constitute a finalized intercomparison protocol and does not directly evaluate full process-based ESM dust schemes. Instead, it provides an observation-constrained pilot framework that can help define where such models should be tested more systematically.

Several limitations follow from this scope. First, the benchmark is based on an interpretable machine-learning reconstruction framework, not on direct evaluation of process-based dust schemes within Earth system models. The present results can therefore identify useful diagnostic dimensions, but they cannot by themselves demonstrate the exact source of structural error in full ESM dust formulations. Second, the validation design is primarily temporal rather than fully spatially independent. This means that reported skill may still be influenced by spatial autocorrelation. Third, the lagged relationships reported here may partly reflect seasonality or temporal autocorrelation in addition to physically meaningful surface persistence. For this reason, the lag diagnostics should be interpreted as benchmark descriptors rather than as stand-alone evidence of mechanism.

A further limitation concerns severe dust events. The present study shows that surface and memory predictors improve overall reconstruction but do not improve extreme-event detection. This means that a climate–surface–memory framework is necessary for a more complete benchmark, but it is not sufficient to solve the tail problem. Future progress will likely require tail-sensitive methods, stricter validation, and data structures that better resolve submonthly and mesoscale processes.

Despite these limitations, the main value of the present study is clear. It shows that dust-model evaluation can be organized more effectively when climate forcing, instantaneous surface state, antecedent surface memory, regime dependence, and tail behaviour are treated as linked but distinct diagnostic dimensions. In that sense, the study provides a practical and reproducible first step toward DesertMIP.

5 Conclusion

This study examined whether dynamic land-surface information and antecedent surface-

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

memory terms improve the reconstruction of dust AOD over North Africa. Three hierarchical configurations were evaluated: Climate-only, Climate plus Surface, and Climate plus Surface plus Memory. The results showed that climate variables alone did not capture the full variability of dust AOD. Adding instantaneous surface predictors improved reconstruction skill, and adding lagged surface information provided a further, although modest, gain.

The memory-sensitive results should be interpreted with care. They show that antecedent surface conditions contain useful information for dust reconstruction, but they do not prove a single causal memory mechanism or a uniquely optimal lag. Instead, they support the use of lag-sensitive diagnostics as part of a broader climate–surface–memory benchmark structure.

The results also showed that model performance was strongly regime dependent. Transition zones were the most difficult benchmark regions, with higher uncertainty and stronger sensitivity to surface and memory information. More homogeneous desert interiors showed smaller gains, while vegetated and semi-vegetated areas benefited more clearly from dynamic surface variables. This finding highlights the need for regime-aware evaluation in dust-model assessment.

Severe dust outbreaks remained difficult to reconstruct. Although surface and memory predictors improved overall reconstruction skill, they did not improve extreme-event detection metrics. Tail behavior therefore remains a separate challenge and should be evaluated with dedicated diagnostics rather than inferred from average model performance.

Taken together, the findings do not establish a finalized community intercomparison protocol. They do, however, provide a reproducible pilot architecture with clearly defined experiments, diagnostics, and benchmark products. The main contribution of the study is the definition of a climate–surface–memory benchmark logic that can guide future evaluation of dust models in dynamic desert environments.

Code and data availability

The DesertMIP project page is available at https://earthsystemsci.github.io/DesertMIP/. The model code used in this study is publicly available through Zenodo at https://doi.org/10.5281/zenodo.20744633 (Mokarram, 2026a). The processed datasets and benchmark masks used to generate the results reported in this manuscript are publicly available through Zenodo at https://doi.org/10.5281/zenodo.19727406 (Mokarram, 2026b). The code archive provides the publicly accessible DesertMIP-CSM v1.0 model code used in this study, and the data archive provides the processed datasets, benchmark masks, and derived outputs required to reproduce the reported results.

Acknowledgements

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

The authors acknowledge the institutional and research support provided by Anhui Xinhua University, Shiraz University, and Nanjing University. The authors also acknowledge Nanjing University for supporting the article processing charge.

Author contributions

Marzieh Mokarram designed the study, developed the machine-learning framework, conducted the analysis, and wrote the first draft of the manuscript. **Mohammad Jafar Mokarram** contributed to data preparation, methodological discussion, and manuscript revision. **Huayu Lu** contributed to the conceptual design of the research, supervised the scientific methodology, and reviewed the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Disclaimer

The publisher remains neutral with regard to jurisdictional claims in published maps, institutional affiliations, and geographical representations included in this article. The responsibility for the content of the article rests entirely with the authors.

Financial support

This work was supported by the National Natural Science Foundation of China under Grant No. 2025zrzdi07.

References

Adebiyi, A.A., Huang, Y., Samset, B.H., Kok, J.F., 2023. Observations suggest that North African dust absorbs less solar radiation than models estimate. Commun. Earth Environ. 4, 168.https://doi.org/10.1029/2021JF006073

Aryal, Y.N., Evans, S., 2021. Global dust variability explained by drought sensitivity in CMIP6 models. J. Geophys. Res. Earth Surf. 126, e2021JF00607. https://doi.org/10.1038/s43247-023-00825-2

Bergametti, G., Marticorena, B., Rajot, J.-L., Chatenet, B., Féron, A., Gaimoz, C., Siour, G., Coulibaly, M., Koné, I., Maman, A., 2017. Dust uplift potential in the central Sahel: An analysis based on 10 years of meteorological measurements at high temporal resolution. J. Geophys. Res. Atmos. 122, 12–433. https://doi.org/10.1002/2017JD027471

Chappell, A., Hennen, M., Schepanski, K., Dhital, S., Tong, D., 2024. Reducing resolution dependency of dust emission modeling using albedo-based wind friction. Geophys. Res. Lett. 51, e2023GL106540. https://doi.org/10.1029/2023GL106540

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E., 2016.

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev. 9, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Ginoux, P., Prospero, J.M., Gill, T.E., Hsu, N.C., Zhao, M., 2012. Global-scale attribution of anthropogenic and natural dust sources and their emission rates based on MODIS Deep Blue aerosol products. Rev. Geophys. 50.

Green, J.K., Berry, J., Ciais, P., Zhang, Y., Gentine, P., 2020. Amazon rainforest photosynthesis increases in response to atmospheric dryness. Sci. Adv. 6, eabb7232.

Huneeus, N., Schulz, M., Balkanski, Y., Griesfeller, J., Prospero, J., Kinne, S., Bauer, S., Boucher, O., Chin, M., Dentener, F., 2011. Global dust model intercomparison in AeroCom phase I. Atmos. Chem. Phys. 11, 7781–7816.

Kim, D., Chin, M., Schuster, G., Yu, H., Takemura, T., Tuccella, P., Ginoux, P., Liu, X., Shi, Y., Matsui, H., 2024. Where dust comes from: Global assessment of dust source attributions with Aerocom models. J. Geophys. Res. Atmos. 129, e2024JD041377.

Kok, J.F., Adebiyi, A.A., Albani, S., Balkanski, Y., Checa-Garcia, R., Chin, M., Colarco, P.R., Hamilton, D.S., Huang, Y., Ito, A., 2021. Contribution of the world's main dust source regions to the global cycle of desert dust. Atmos. Chem. Phys. 21, 8169–8193.

Kok, J.F., Storelvmo, T., Karydis, V.A., Adebiyi, A.A., Mahowald, N.M., Evan, A.T., He, C., Leung, D.M., 2023. Mineral dust aerosol impacts on global climate and climate change. Nat. Rev. Earth Environ. 4, 71–86.

LeGrand, S.L., Letcher, T.W., Okin, G.S., Webb, N.P., Gallagher, A.R., Dhital, S., Hodgdon, T.S., Ziegler, N.P., Michaels, M.L., 2023. Application of a satellite-retrieved sheltering parameterization (**v1.0**) for dust event simulation with WRF-Chem **v4.1**. Geosci. Model Dev. 16, 1009–1038.

Mahowald, N., Albani, S., Kok, J.F., Engelstaeder, S., Scanza, R., Ward, D.S., Flanner, M.G., 2014. The size distribution of desert dust aerosols and its impact on the Earth system. Aeolian Res. 15, 53–71.

Mokarram, M., 2026a. DesertMIP-CSM v1.0 model code. Zenodo. https://doi.org/10.5281/zenodo.20744633.

Mokarram, M., 2026b. DesertMIP processed datasets and benchmark masks, version 1 [data set]. Zenodo. https://doi.org/10.5281/zenodo.19727406

Molnar, C., 2020. *Interpretable machine learning: A guide for making black box models explainable*. Lulu.com.https://christophm.github.io/interpretable-ml-book/

Mousavi, S.V., Karami, K., Tilmes, S., Muri, H., Xia, L., Rezaei, A., 2023. Future dust concentration over the Middle East and North Africa region under global warming and stratospheric aerosol intervention scenarios. Atmos. Chem. Phys. 23, 10677–10695.

Pu, B., Ginoux, P., 2018. How reliable are CMIP5 models in simulating dust optical depth?

https://doi.org/10.5194/egusphere-2026-2271
Preprint. Discussion started: 19 June 2026
© Author(s) 2026. CC BY 4.0 License.

Atmos. Chem. Phys. 18, 12491–12510.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, F., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.

Ridley, D.A., Heald, C.L., Prospero, J.M., 2014. What controls the recent changes in African mineral dust aerosol across the Atlantic? Atmos. Chem. Phys. 14, 5735–5747. https://doi.org/10.5194/acp-14-5735-2014

Rohrmann, A., Heermance, R., Kapp, P., Cai, F., 2013. Wind as the primary driver of erosion in the Qaidam Basin, China. Earth Planet. Sci. Lett. 374, 1–10.

Soci, C., Hersbach, H., Simmons, A., Poli, P., Bell, B., Berrisford, P., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Radu, R., 2024. The ERA5 global reanalysis from 1940 to 2022. Q. J. R. Meteorol. Soc. 150, 4014–4048.

Van De Kerchove, R., Zanaga, D., Keersmaecker, W., Souverijns, N., Wevers, J., Brockmann, C., Grosu, A., Paccini, A., Cartus, O., Santoro, M., 2021. ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data., in: AGU Fall Meeting Abstracts. pp. GC45I-0915.

Wang, W., Evan, A.T., Flamant, C., Lavaysse, C., 2015. On the decadal scale correlation between African dust and Sahel rainfall: The role of Saharan heat low–forced winds. Sci. Adv. 1, e1500646.

Wen, J., Tagliabue, G., Rossini, M., Fava, F. Pietro, Panigada, C., Merbold, L., Leitner, S., Sun, Y., 2025. Detection of fast-changing intra-seasonal vegetation dynamics of drylands using solar-induced chlorophyll fluorescence (SIF). Biogeosciences 22, 2049–2067.

Wen, Q., Li, Y., Du, M., Song, W., Wei, L., Wang, Z., Li, X., 2025. Interdecadal shift in the impact of winter land–sea thermal contrasts on following spring transcontinental dust transport pathways in North Africa. Atmos. Chem. Phys. 25, 10853–10867.

Wu, C., Lin, Z., He, J., Zhang, M., Liu, X., Zhang, R., Brown, H., 2016. A process-oriented evaluation of dust emission parameterizations in CESM: Simulation of a typical severe dust storm in E ast A sia. J. Adv. Model. Earth Syst. 8, 1432–1452.

Zhang, B., Xiong, D., Tang, Y., Liu, L., 2022. Land surface roughness impacted by typical vegetation restoration projects on aeolian sandy lands in the Yarlung Zangbo River valley, southern Tibetan plateau. Int. Soil Water Conserv. Res. 10, 109–118.

Zhang, L., Montuoro, R., McKeen, S.A., Baker, B., Bhattacharjee, P.S., Grell, G.A., Henderson, J., Pan, L., Frost, G.J., McQueen, J., 2022. Development and evaluation of the aerosol forecast member in the National Center for Environment Prediction (NCEP)'s global ensemble forecast system (GEFS-Aerosols v1). Geosci. Model Dev. 15, 5337–5369.