Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 1, 2026

Environmental Modulation of Interpolation Strategies and Algorithm Adaptability in Remote Sensing Retrieval of Farmland Soil Moisture: A Theoretical Framework

Dongqiang Yang *

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China

Abstract: Accurate retrieval of farmland soil moisture from multi-source remote sensing data requires simultaneous consideration of time-series interpolation strategies and machine learning algorithm selection. However, the theoretical basis explaining how environmental conditions modulate their joint effects remains insufficiently articulated. This paper proposes a unified conceptual framework termed the "Interpolation-Algorithm-Environment" (IAE) triad, which explains why no universally optimal retrieval pipeline exists and why optimal choices are inherently environment-contingent. We further introduce the concept of Post-Treatment Spurious Correlation (PTSC) to characterize a failure mode in which interpolation processing inadvertently introduces artificial feature structure that retrieval models learn and amplify, degrading generalization. Theoretical analysis of algorithm-environment matching, interpolation-induced information distortion, and synergistic constraint mechanisms provides a transferable decision basis for designing robust soil moisture retrieval systems under diverse agro-climatic conditions.

Keywords: Soil moisture retrieval, Time-series interpolation, Machine Learning.

1. Introduction

Farmland soil moisture is one of the most operationally significant geophysical variables in the Earth system, mediating precipitation partitioning, plant-available water, crop productivity, and regional land-surface energy exchange. High-spatiotemporal-resolution soil moisture monitoring is therefore a prerequisite for precision agriculture, drought early warning, and sustainable water resource management.

Microwave remote sensing—particularly synthetic aperture radar (SAR) at meter-to-decameter resolution—provides the physical basis for operational soil moisture retrieval via the dielectric contrast between moist and dry soils. When complemented by optical vegetation indices (e.g., NDVI, NDWI) from platforms such as Sentinel-2, SAR-based retrieval can partially decouple the soil moisture signal from vegetation and roughness confounders. This multi-source fusion paradigm has become dominant for achieving broad spatial coverage and fine resolution [1].

A pervasive challenge undermines the consistency of this approach: optical time-series data are inherently discontinuous owing to cloud cover, atmospheric contamination, and sensor scheduling gaps—especially in humid and monsoon-affected landscapes [2]. The interpolation method used to fill these gaps shapes the covariance structure of the feature space presented to the learner, potentially creating spurious correlations that models exploit during training but that do not generalize to new sites or time periods [3].

Despite the practical importance of this coupling, existing literature treats interpolation and algorithm selection as largely separable concerns, missing the crucial interaction between feature generation and model induction. The present paper argues that interpolation strategy (I), retrieval algorithm (A), and environmental context (E) form an irreducible triad, and that a theoretically grounded framework for their joint dynamics is both scientifically necessary and practically valuable.

2. Environmental Regulation of Remote Sensing Signal Information Content

2.1. The Multi-Source Signal as an Environmentally Mediated Composite

The information carried by a multi-source remote sensing feature vector is jointly determined by sensor characteristics, atmospheric state, land surface conditions, and the temporal sampling interval. This environmental mediation operates through three primary pathways.

The soil moisture-backscatter relationship is itself environmentally contingent. In humid subtropical environments with dense, rapidly changing canopies, vegetation may dominate the backscatter during certain phenological stages, compressing the dynamic range of the soil moisture signal. In arid or semi-arid environments, surface roughness introduced by tillage may become the dominant backscatter driver. The slope, offset, and nonlinearity of the soil moisture-feature relationship are therefore functions of the local environment, not universal constants.

Second, the temporal autocorrelation structure of soil moisture varies systematically across climatic regimes. Precipitation-dominated humid systems exhibit strong event-driven dynamics—rapid wetting after rainfall followed by exponential drainage—with characteristic timescales of days to weeks. In terrain-controlled systems, temporal variability is lower but spatial heterogeneity is high and structured by landscape position. These contrasting dynamics directly determine which machine learning architectures can best

exploit the available information.

Third, optical data missing patterns are not environmentally random. Cloud cover correlates with rainfall and hence with high soil moisture; systematic removal of cloudy observations introduces conditional bias in the observed feature distribution. Any interpolation method that neglects this bias will systematically underestimate optical indices during periods of actual high soil moisture, cascading into retrieval error.

2.2. Effective Information Bandwidth

We introduce the concept of effective information bandwidth (EIB) to characterize the degree to which a given multi-source feature set can resolve soil moisture variation under specific environmental conditions. EIB is conceptually defined as the mutual information between the feature vector and the target variable, integrated over the relevant temporal and spatial scales. High EIB indicates a strong, unambiguous soil moisture signal; low EIB indicates compression by confounders or data quality limitations.

EIB decreases monotonically with increasing optical data missing rate under any interpolation strategy that does not recover lost information from independent data sources. Among univariate methods, those preserving underlying temporal trends (e.g., polynomial smoothing) maintain higher EIB than those merely enforcing temporal continuity. Multivariate interpolation methods incorporating SAR backscatter as an auxiliary variable can partially recover EIB by constraining reconstructions to be physically consistent with observed microwave signals—motivating the practical finding that SAR-informed interpolation provides more stable cross-site generalization than purely temporal univariate methods under concentrated optical data loss.

3. Interpolation-Induced Information Distortion and Model Learning

3.1. Structural Distortion of the Feature Space

Any interpolation method transforms an observed (incomplete) feature time series into a reconstructed (complete) one, necessarily introducing assumptions about unobserved values. When these assumptions are satisfied—e.g., when the underlying process is smooth and missing patterns are genuinely random—interpolation improves signal-to-noise ratio without systematic error. When assumptions are violated, as in the cloud-correlated missing scenario, interpolation distorts the covariance structure of the feature space.

The case of high missing rates with concentrated temporal gaps is particularly critical. In this regime, univariate methods extrapolate beyond observed data using temporally distant and climatically distinct observations, producing smooth, low-frequency interpolated segments that are statistically distinguishable from genuine observations. This bimodal feature distribution can confound model training when the model has sufficient capacity to detect and exploit this distributional signature.

The severity of structural distortion scales with three factors: (1) the missing rate—higher rates produce longer interpolated segments with lower fidelity; (2) the interpolation method—stronger smoothness priors produce more extreme low-frequency distortion; (3) the environmental dynamics—environments with rapid soil moisture transitions are more sensitive to distortion because the temporal gradient of the true signal is steeper.

3.2. Post-Treatment Spurious Correlation (PTSC)

We define Post-Treatment Spurious Correlation (PTSC) as a statistical artifact arising when a retrieval model learns associations between interpolation-induced feature structure and the target variable that do not reflect genuine physical relationships and therefore do not generalize beyond the training distribution. PTSC is distinguished from ordinary overfitting in that it is specifically caused by the preprocessing treatment rather than by model complexity per se.

PTSC can be operationally detected through three diagnostic criteria. (1) Error-feature correlation: systematic correlation of model residuals with interpolated feature values in genuinely missing segments indicates that the model has learned a bias linked to interpolation artifacts. (2) Smoothness dependence: monotonic performance degradation as greater smoothing is applied to input features suggests the model's advantage in the unsmoothed case was partly attributable to high-frequency feature variation from interpolation artifacts. (3) Generalization discrepancy: a disproportionately large performance gap between spatially random cross-validation and leave-one-site-out validation indicates that cross-validation performance was inflated by PTSC.

PTSC is not uniformly distributed across algorithm classes. LSTM and other recurrent models are more susceptible than ensemble tree methods because the memory state provides a pathway through which smooth interpolated segments propagate and amplify temporally. This temporal amplification mechanism does not exist in ensemble tree methods, which make predictions based on local feature values without explicit sequence state. Consequently, LSTM's performance advantage over RF attenuates and eventually reverses as missing rates increase.

3.3. The Critical Missing Rate Threshold

PTSC severity as a function of missing rate suggests a critical threshold above which risk changes qualitatively from manageable to severe. Below this threshold, interpolated segments are short relative to the soil moisture temporal autocorrelation length, and interpolated values are tolerably close to true values. Above the threshold, interpolated segments extend beyond the autocorrelation length, increasingly reflecting the interpolation method's prior assumptions rather than the actual environmental state.

The location of this threshold is environmentally determined: in environments with long soil moisture memory (slow-draining soils, stable topographic control), the threshold is higher; in environments with rapid dynamics (episodic precipitation, sandy soils with rapid drainage), the threshold is lower. This motivates matching the interpolation method to the characteristic temporal dynamics of the study environment.

4. Algorithm-Environment Matching

4.1. Information Structure Compatibility

Machine learning retrieval models differ fundamentally in the type of information structure they exploit. Temporal deep learning architectures (LSTM, TCN) carry an inductive bias that the order and timing of feature observations carry information beyond contemporaneous feature values—

appropriate when soil moisture exhibits genuine temporal dependencies characteristic of precipitation-driven systems. Ensemble tree methods (RF, XGBoost) assume the target variable is predictable from a function of contemporaneous feature values, making each observation independent given its features—better suited to environments where static or slowly varying factors explain most spatial variability in soil moisture.

This compatibility argument generates a falsifiable prediction: algorithm performance advantage should correlate positively with the degree to which the environment's information structure matches the algorithm's inductive bias. In precipitation-driven systems, LSTM should outperform RF [5]; in terrain-controlled systems, RF should outperform LSTM—providing a theoretical basis for observed regional specificity in optimal algorithm selection.

4.2. Robustness Under Data Quality Perturbation

Ensemble methods achieve robustness through two structural mechanisms: aggregating predictions across many trees trained on bootstrap samples averages out the influence of individual noisy observations, and randomized feature subsampling in RF prevents any single noisy feature from dominating predictions across all trees. These mechanisms provide structural resistance to localized feature perturbations introduced by interpolation.

LSTM lacks inherent mechanisms for averaging out sequential noise: a smooth interpolated segment that passes through the network at training time updates weights cumulatively, amplified by the length and frequency of interpolated segments. This structural asymmetry explains why LSTM's performance advantage over RF is not constant across missing rate conditions: at low missing rates, LSTM's temporal modeling advantage dominates; at high missing rates, RF's robustness advantage dominates.

4.3. The Algorithm Selection Decision Boundary

Combining the information structure compatibility argument with the robustness argument, algorithm selection can be characterized in a two-dimensional space defined by (1) the strength of temporal dependencies in the target environment and (2) the optical data missing rate. LSTM dominates when temporal dependencies are strong and missing rates are low; RF dominates when temporal dependencies are weak or missing rates are high. The decision boundary is a gradient rather than a sharp line, modulated by specific environmental dynamics and data characteristics.

For practitioners, this means first characterizing the temporal autocorrelation structure of soil moisture from available in-situ data, and estimating the expected optical data missing rate from historical cloud cover statistics. These two quantities together determine which algorithm class is likely to be optimal prior to any hyperparameter tuning.

5. The IAE Triad Framework

5.1. Formal Structure

The IAE triad framework formalizes the joint determination of retrieval performance by three interacting factors. Let I denote the interpolation strategy, A the retrieval algorithm, and E the environmental context (climate type, terrain complexity, temporal soil moisture dynamics, optical data missing rate). The retrieval performance metric R(I, A, E) includes substantial interaction terms:

$$R(I, A, E) = \mu + \alpha(I) + \beta(A) + \gamma(E) + (I\times A) + (I\times E) + (A\times E) + (I\times A\times E)$$

The central claim is that the three-way interaction term (I x A x E) is non-negligible—it may dominate performance variance across deployment contexts—and that retrieval pipeline designs ignoring it will yield suboptimal, context-dependent results. Performance evaluations conducted in a single environment with a single interpolation strategy do not generalize; a full-factorial experiment across multiple environments and preprocessing strategies is needed to directly estimate the interaction structure of the triad.

5.2. Synergistic Constraint

The IAE triad exhibits a synergistic constraint structure: the performance ceiling achievable by any (I, A) combination is bounded by the EIB of the environment, which is partly determined by interpolation quality. This creates a cascading dependency: environmental conditions set the upper boundon EIB; interpolation quality determines how much of that EIB is preserved in the feature representation; algorithm architecture determines how much of the preserved EIB is exploited for prediction.

An important corollary is that improving any single component yields diminishing returns if other components are mismatched. Conversely, optimizing all three components in alignment yields compounding benefits: physically appropriate interpolation preserves high EIB, which is efficiently exploited by a matched algorithm architecture.

5.3. Decision Heuristics for Practitioners

The IAE framework generates four practitioner heuristics. First, characterize the environment before selecting methods: estimate the temporal autocorrelation length of soil moisture, the expected optical data missing rate, and the degree of terrain control on spatial soil moisture variation. These diagnostics determine the optimal region of the (I, A) design space.

Second, match interpolation method to missing pattem type: when missing data are scattered and individually short (days to a week), univariate smoothing methods such as Savitzky-Golay filtering preserve trend fidelity with low computational cost [4]. When missing data are concentrated in long continuous gaps (multiple weeks), multivariate methods incorporating SAR data—such as MSSA or ARIMAX—provide more reliable reconstruction.

Third, match algorithm architecture to environmental dynamics: in precipitation-dominated environments with strong temporal autocorrelation, LSTM offers superior performance when data quality is adequate. In terrain-controlled or seasonally stable environments, or when optical data missing rates exceed approximately 60%, RF and XGBoost provide better robustness and generalization with lower PTSC sensitivity.

Fourth, diagnose PTSC before deploying: apply the three diagnostic criteria (error-feature correlation, smoothness dependence, generalization discrepancy) to the training pipeline before operational deployment. If PTSC is detected, consider switching to a more PTSC-resistant algorithm, reducing reliance on heavily interpolated features, or prioritizing improvements to raw data quality over further interpolation refinement.

6. Conclusions

This paper proposes a fundamental reconceptualization of the farmland soil moisture retrieval problem as an irreducible triad encompassing interpolation strategy, retrieval algorithm, and environmental context. The principal theoretical contributions are as follows.

First, environmental conditions govern the effective information bandwidth (EIB) of multi-source remote sensing features through three pathways: modulation of the soil moisture-backscatter relationship, shaping of temporal autocorrelation structure, and conditioning of optical data missing patterns. These pathways collectively determine the maximum achievable retrieval performance for any pipeline operating within a specific environment.

Second, Post-Treatment Spurious Correlation (PTSC) is formalized as a specific failure mode wherein interpolation-induced feature structure is learned by retrieval models, degrading generalization. Three operational diagnostic criteria for PTSC detection are characterized, and architectural properties are identified that render temporal sequence models (LSTM) more prone to PTSC than ensemble tree methods (RF, XGBoost), especially at high optical data missing rates.

Third, an algorithm-environment matching principle is derived from the compatibility between a model's inductive bias (temporal sequence modeling versus contemporaneous feature mapping) and the information structure of the target environment (dynamic temporal dependencies versus static spatial control). This principle offers a theoretical foundation for the empirical pattern that LSTM outperforms RF in precipitation-dominated systems, while RF outperforms LSTM in terrain-controlled systems.

Fourth, the IAE triad framework formalizes the joint determination of retrieval performance by all three factors and their interactions, generating practical decision heuristics for interpolation strategy and algorithm selection contingent on environmental characterization, and prompting a reframing of retrieval evaluation protocols to explicitly estimate interaction effects across preprocessing strategies and environmental contexts.

Collectively, these contributions provide a transferable theoretical basis for designing robust, context-aware soil moisture retrieval pipelines—a fundamental prerequisite for improving the precision, scalability, and operational reliability of remote-sensing-based farmland monitoring.

Acknowledgements

The authors declare that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Ulaby, F.T., Moore, R.K., & Fung, A.K. (1978). Microwave remote sensing active and passive. Addison-Wesley.

[2] Njoku, E.G., & Kong, J.A. (1977). Theory for passive microwave remote sensing of near-surface soil moisture. Journal of Geophysical Research, 82(20), 3108-3118.

[3] Bindlish, R., & Barros, A.P. (2000). Multifrequency soil moisture inversion from SAR measurements with the use of IEM. Remote Sensing of Environment, 71(1), 67-88.

[4] Savitzky, A., & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8), 1627-1639.

[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.