*remote sensing*

MDPI

*Article*

# Diagnosing the Applicability Boundary of Water Cloud Model-Based Soil Moisture Retrieval Under Cross-Site Generalization

**Peng Wang, Changbao Yang and Qigang Jiang ***

College of Geoexploration Science and Technology, Jilin University, Changchun 130026, China
* Correspondence: jiangqigang@jlu.edu.cn

## Highlights

### What are the main findings?

- Leave-one-station-out (LOSO) $R^2$ is near zero at all four networks, with or without Water Cloud Model (WCM) features.
- LOSO skill decreases with landscape heterogeneity; anomaly gains are largely leakage.

### What is the implication of the main finding?

- Random-split scores overstate the cross-site skill of WCM-ML retrieval.
- A brief in situ warm-up (~10% of record) corrects the shift at heterogeneous sites.

## Abstract

The Water Cloud Model (WCM) is the most widely used vegetation-scattering correction framework for microwave soil-moisture (SM) remote sensing, yet its cross-site generalization ability remains poorly understood. This study compares WCM-based random forest models across four validation networks spanning a landscape-heterogeneity gradient (REMEDHUS, SMOSMANIA, Little Washita, and OzNet; station-mean coefficient of variation (CV) of 0.18–0.67), under both within-station random-split and leave-one-station-out (LOSO) validation. The WCM-decoupled feature yields no significant accuracy gain under random validation (Wilcoxon $p \geq 0.168$), and all models fall below a monthly climatology baseline under LOSO, consistent with a generalization failure that is more plausibly associated with station-level distributional shift than with vegetation-correction error. LOSO $R^2$ decreases monotonically with landscape heterogeneity across the four networks, but at the subnetwork level it is robust only at the most heterogeneous network (REMEDHUS); in homogeneous landscapes, temporal transferability appears more important. A data-leakage artifact is identified and corrected in anomaly-based modeling, substantially reducing apparent gains at two of the four sites. A target-station-mean correction calibrated from only the earliest 10% of a new station's observations lifts $R^2$ above the climatology baseline at the three most heterogeneous networks, suggesting that much of the failure is a correctable station-baseline shift, although other factors cannot be excluded. These findings provide supporting evidence on the applicability boundary of WCM–machine-learning frameworks for network-scale soil moisture retrieval, based on the four networks examined.

**Keywords:** soil moisture; water cloud model; cross-site generalization; landscape heterogeneity; data leakage; domain correction

Check for updates

Academic Editor: Gabriel Senay

Received: 15 September 2026
Revised: 23 October 2026
Accepted: 6 November 2026
Published: 8 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Soil moisture (SM) is a key state variable linking atmospheric, land-surface, and groundwater processes [1], and is the core observable of the first dedicated satellite missions: the Soil Moisture and Ocean Salinity (SMOS) mission, launched by the European Space Agency (ESA) in 2009 [2], and the Soil Moisture Active Passive (SMAP) mission, launched by NASA in 2015 [3]. The failure of the SMAP radar in July 2015 limited its soil moisture retrieval to a resolution of about 40 km rather than the originally planned 9 km; this limitation has been partly mitigated by fusing SMAP and SMOS observations [4]. Sentinel-1 C-band synthetic aperture radar (SAR), with its high spatial resolution and all-weather capability, has become an important source of SM information [5], although canopy scattering limits retrieval accuracy. Ground validation relies mainly on networks such as REMEDHUS [6], Little Washita [7], OzNet [8], and SMOSMANIA [9], together with their international integration platform, the International Soil Moisture Network (ISMN) [10].

The Water Cloud Model (WCM) [11] offers a physically interpretable way to correct for vegetation, and its C-band parameter calibration has been studied systematically [12–14]. Combining WCM with machine-learning (ML) methods such as random forest (RF) [15] and extreme gradient boosting (XGBoost) [16,17] has achieved, for example, RMSE = 4.0–6.1%vol [18], $R^2$ = 0.49–0.88 [19], and a backscatter RMSE of 1.2–1.3 dB [20] under random or field-level splitting, and an average per-sensor correlation of about 0.73 on held-out sensors [21]. The wider landscape of ML approaches to SM analysis, from time-series to deep-learning and physics-informed methods, has recently been surveyed in detail [22].

This framework builds on a long line of semi-empirical and physical models of rough-surface scattering, including the integral equation model [23] and its calibration refinements based on field SAR data [24], the Oh [25] and Dubois [26] models, C-band roughness parameterizations [27], and change-detection approaches [28]. Its theoretical roots trace back to classical work on passive-microwave vegetation correction [29] and to systematic reference texts [30]. On the platform side, the SMOS mission [2] and the Sentinel-1 mission [31], whose system description underlies the SAR data used in this study, are key components of the SM remote-sensing data landscape, and the L-band NASA–ISRO Synthetic Aperture Radar (NISAR) mission will soon add to them [32].

However, existing research has two shortcomings. First, most studies rely on random or sensor-level split validation, which rarely probes cross-network extrapolation, whereas real-world deployment requires cross-site generalization, evaluated by leave-one-station-out (LOSO) validation. The ecological and geoscience modeling literature has repeatedly shown that random splitting of spatially autocorrelated data overestimates model performance, and that spatial blocking or leave-one-region-out validation should be used instead [33–37]. This concern is not unique to soil moisture: models trained in one region lose substantial accuracy when transferred to another, even within the same remote-sensing domain, as shown for crop classification [38]. It was also evident in our own earlier Sentinel-1 study at Little Washita, where a stacked-ensemble model without lagged soil-moisture inputs reached $R^2$ = 0.63 under a temporal hold-out but a mean $R^2$ of −0.60, negative in all five folds, at stations withheld from training [39]. Second, the modulating effect of landscape type on WCM effectiveness has not been systematically studied. The central question of this study is therefore: how does landscape heterogeneity, indexed by the CV of station-mean SM, affect the cross-site generalization ability of WCM-ML, and under what landscape conditions does this effect dominate? The first law of geography, that nearby observations are more strongly correlated than distant ones [40], helps explain why random splitting of spatially autocorrelated data overestimates cross-site generalization.

To address this question, the study compares four independent validation networks (REMEDHUS in Spain, SMOSMANIA in France, Little Washita in the United States, and OzNet Yanco in Australia), selected on three criteria. First, each provides multi-station, multi-year in situ SM records that overlap the Sentinel-1 era and are publicly available. Second, together they span a wide gradient of landscape heterogeneity, with the coefficient of variation (CV) of station-mean SM ranging from 0.18 (OzNet) to 0.67 (REMEDHUS), so that the effect of heterogeneity can be examined over a continuous range rather than at a single site. Here heterogeneity is operationalized as the CV of station-mean SM, an empirical index that integrates the effects of land cover, soil properties, topography, and local climate on each station's SM level, and is therefore not a pure measure of landscape structure. Third, they include networks that differ in climate but have similar heterogeneity (SMOSMANIA, CV = 0.38; Little Washita, CV = 0.36), which allows the effect of heterogeneity to be partly distinguished from that of climate. Applying the same canonical pipeline across these networks is therefore a main objective of this study rather than a convenience: it tests whether the generalization behavior of WCM-ML observed at one site reproduces under different landscape and climate conditions, and where its applicability boundary lies.

Contributions of this study: (1) a systematic cross-network comparison across four validation networks that span a gradient of landscape heterogeneity (station-mean CV 0.18–0.67) and different climates (including the French SMOSMANIA network, spanning a Mediterranean–Atlantic climate gradient), using a unified canonical pipeline to reduce methodological confounds; (2) identifying and correcting a data-leakage issue in anomaly modeling—such leakage has been shown to be a major source of the reproducibility crisis in machine-learning-driven research [41]; (3) large-scale fixed-n subnetwork analysis (300+ random combinations) to control for sample-size confounds; (4) strengthening the reliability of the core conclusions through canonical-RF hyperparameter sensitivity checks; (5) integrating error decomposition, baseline comparison, and mechanism verification into a diagnostic framework; (6) transparently reporting that the WCM B parameter could not be calibrated to a clear optimum at the OzNet and SMOSMANIA networks (the grid search, expanded to 0.05–20, finds no distinct interior optimum), and showing through a vegetation-term sensitivity test that the main conclusions do not depend on this treatment.

## 2. Materials and Methods

### *2.1. Study Area Overview*

This study selects four internationally recognized soil moisture validation networks that differ substantially in landscape type, vegetation cover, and inter-station heterogeneity, forming an ideal comparative set for analyzing the applicability boundary of WCM (Table 1; Figure 1). The scientific basis of soil moisture's spatial–temporal variability and observation-network design has been the subject of long-standing systematic research [42–46]; the probes at all four sites are based on the classical principle of measuring soil dielectric permittivity via time-domain reflectometry or capacitance methods to retrieve volumetric water content [47]. Global-scale reanalysis and fusion products (such as the Global Land Data Assimilation System (GLDAS) [48], the ESA Climate Change Initiative (CCI) Soil Moisture climate data record [49,50], and the COsmic-ray Soil Moisture Observing System (COSMOS) network [51]) further enrich validation approaches beyond ground networks, but station-scale ground observations remain an irreplaceable benchmark for evaluating these products.

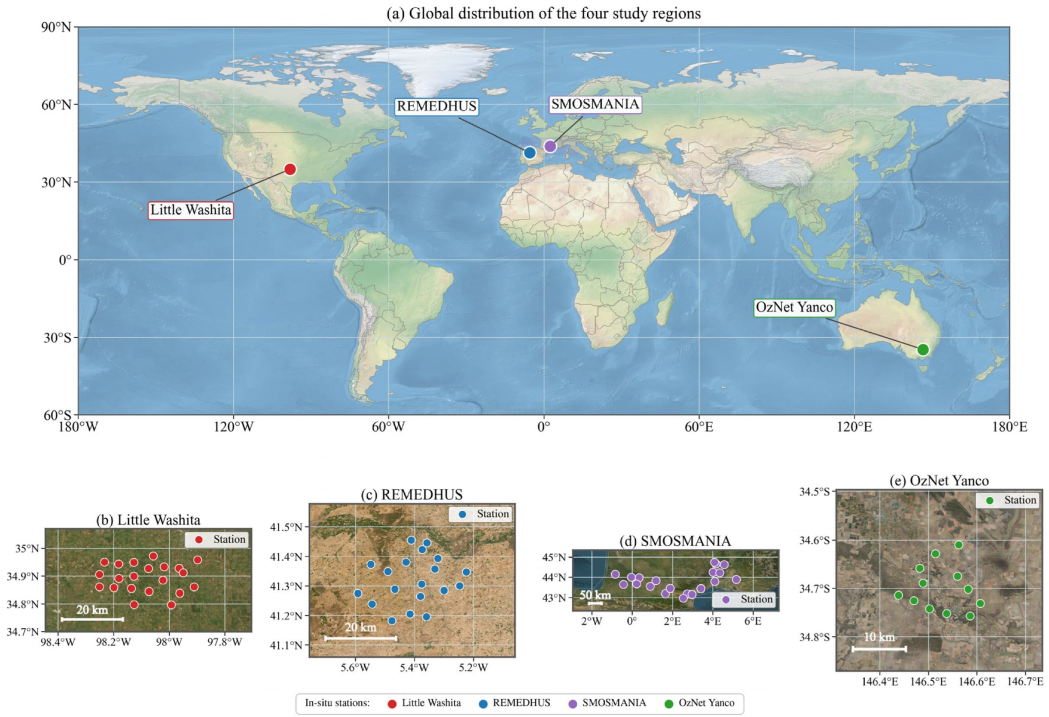



**Figure 1.** Geographic locations of the four soil moisture validation networks used in this study. (**a**) Global distribution of the four study regions; (**b**–**e**) station layouts of Little Washita, REMEDHUS, SMOSMANIA, and OzNet Yanco. Dots denote in situ stations.

**Table 1.** Basic characteristics of the four study areas.

| Feature | REMEDHUS | Little Washita | OzNet Yanco | SMOSMANIA | Unit |
|---|---|---|---|---|---|
| Number of stations | 19 | 21 | 12 | 21 | stations |
| Data period | 2019–2023 | 2016–2023 | 2017–2023 | 2016–2023 | — |
| Landscape type | Heterogeneous rainfed/irrigated agriculture | Semi-homogeneous mixed agro-pastoral | Homogeneous semi-arid grassland | Mediterranean–Atlantic climate-gradient agro-forestry mosaic | — |
| Mean SM CV * | 0.673 | 0.364 | 0.184 | 0.380 | dimensionless |
| WCM B parameter | 1.900 § | 0.900 § | 0.300 † | 0.243 † | — |
| Sample size | 3730 | 9997 | 2108 ‡ | 17,808 | records |

* CV = coefficient of variation in the multi-year mean SM across stations (standard deviation/mean; dimensionless and invariant to the unit convention); it is used as an empirical proxy for heterogeneity and also reflects climatic, soil, and hydrological differences among stations. Station-mean SM is expressed in $m^3\ m^{-3}$ throughout this study (OzNet/SMOSMANIA divided by 100 from their native %vol scale; REMEDHUS/Little Washita are native $m^3\ m^{-3}$ and unchanged), consistent with Figure 2a. † The WCM B parameters for OzNet (B = 0.300) and SMOSMANIA (B = 0.243) are literature defaults rather than calibrated values—the grid search (including an expanded re-check over 0.05–20) did not identify a clear interior optimum at either network; see Sections 2.2.3 and 4.5 for details. ‡ The OzNet sample size (Figure 2c) is the number of records after cleaning (2108; see Supplementary Section S1), consistent with the analyses in Sections 3.2–3.6. § REMEDHUS and Little Washita B values were confirmed via an expanded grid re-check (0.05–20; Supplementary Section S2): REMEDHUS B = 1.900 and Little Washita B = 0.900. REMEDHUS additionally uses region-specific vegetation parameters A_VV = 0.0046, A_VH = 0.0017 (Baghdadi et al. [12], calibrated for winter-cereal canopies) rather than the A_VV = 0.10, A_VH = 0.05 shared by the other three regions; see Section 2.2.2 for detail.

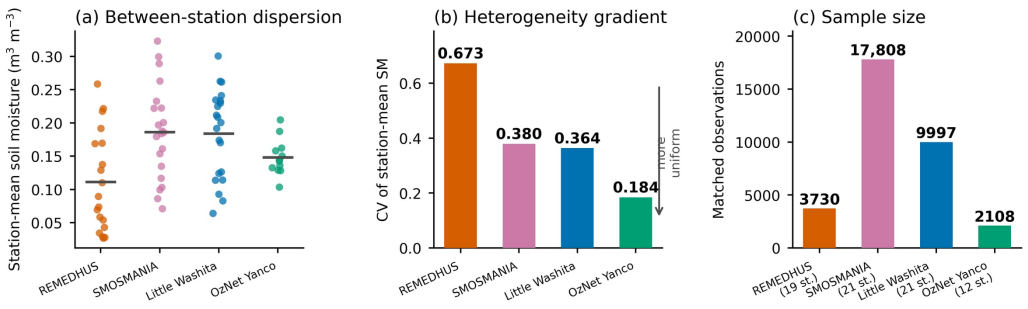



**Figure 2.** Landscape heterogeneity gradient across the four validation regions. (**a**) Distribution of multi-year mean soil moisture at each station, in $m^3$ $m^{-3}$ throughout (OzNet/SMOSMANIA divided by 100 from their native %vol scale for consistency with the native-fraction REMEDHUS/Little Washita convention), with horizontal lines indicating regional means; (**b**) heterogeneity gradient formed by the station-mean coefficient of variation (CV); (**c**) matched observation sample size and station count. CV values for the four regions are 0.673 (REMEDHUS), 0.380 (SMOSMANIA), 0.364 (Little Washita), and 0.184 (OzNet Yanco). SM, soil moisture; CV, coefficient of variation; dots in (**a**) denote individual stations and horizontal lines the regional means; "st." in (**c**) denotes stations.

The observation periods are 5–8 years for the four networks (REMEDHUS, 2019–2023; OzNet, 2017–2023; Little Washita and SMOSMANIA, 2016–2023; Table 1). All four networks share the common window 2019–2023, and each period spans at least five full annual cycles of the Sentinel-1 record, sampling multiple wet–dry cycles and interannual variability. The resulting matched samples (3730 at REMEDHUS, 2108 at OzNet, 9997 at Little Washita, and 17,808 at SMOSMANIA, from 12 to 21 stations each) provide sufficient data for leave-one-station-out validation, which is a spatial rather than a temporal comparison.

(1) **REMEDHUS (Spain)**

The REMEDHUS network is located in Salamanca Province, northwestern Spain, and comprises 19 stations equipped with 5TM sensors, covering approximately 1300 $km^2$ of agricultural land in the Duero basin. Land use is highly heterogeneous—rainfed cereal, irrigated agriculture, woodland, and grassland are interspersed—resulting in substantial inter-station differences in mean SM (range 0.027–0.259 $m^3$ $m^{-3}$, CV = 0.673). Vegetation cover is dominated by winter wheat, sunflower, and barley, with strong seasonal Normalized Difference Vegetation Index (NDVI) variation [6].

(2) **Little Washita (USA)**

The Little Washita micro-watershed is located in southwestern Oklahoma, USA, and comprises 21 stations equipped with Hydra Probe sensors, covering approximately 611 $km^2$ of semi-arid mixed agriculture-grassland. Land use is dominated by winter wheat and grassland; inter-station SM differences are lower than at REMEDHUS (CV = 0.364), but notable spatiotemporal heterogeneity remains [7].

(3) **OzNet Yanco (Australia)**

The OzNet Yanco network (hereafter OzNet) is located in the Murrumbidgee catchment, New South Wales, Australia, and comprises 12 stations covering approximately 2200 $km^2$ of homogeneous grassland. Land use is dominated by natural grassland and sparse shrubland, with low and spatially uniform vegetation cover, resulting in the smallest inter-station SM difference among the four regions (CV = 0.184) [8].

(4) **SMOSMANIA (France)**

The SMOSMANIA network was established by Météo-France by adding soil moisture probes to its RADOME automatic weather station network, and became operational in Jan-

uary 2007 [9]. The network comprises 21 stations distributed linearly along a Mediterranean–Atlantic climate gradient in southwestern France, with an average inter-station spacing of about 40 km, spanning a pronounced precipitation and vegetation gradient (from Mediterranean arid shrubland to Atlantic temperate agro-forestry mosaics). Each station is equipped with ThetaProbe ML2X capacitance probes at four depths (5/10/20/30 cm); this study uses the shallowest layer (5 cm) to match SAR penetration depth. The inter-station SM CV (0.380) is close to that of Little Washita (0.364), but is jointly driven by two markedly different climate regimes, providing an independent cross-climate sample for testing whether CV alone fully characterizes cross-site generalization ability; see Sections 4.3 and 4.7 for discussion.

### *2.2. Data Preprocessing and WCM Parameter Calibration*

#### 2.2.1. Data Processing

SAR data for all four regions were taken from Sentinel-1 Ground Range Detected products (Interferometric Wide mode; VV + VH dual polarization, i.e., vertical transmit–vertical receive and vertical transmit–horizontal receive; ascending and descending orbits where available; the OzNet data are descending-only). VV and VH backscatter coefficients (dB) and per-scene incidence angles were extracted at the station locations via Google Earth Engine (extraction settings are given in Supplementary Section S1). NDVI [52,53] was derived from Sentinel-2 surface reflectance, also accessed via Google Earth Engine [54], and matched to each Sentinel-1 observation in time (nearest valid scene, or temporal interpolation at SMOSMANIA; matching windows are given in Supplementary Section S1). Cloud-screening thresholds and extraction settings are given in Supplementary Section S1.

At REMEDHUS, Little Washita, and OzNet, Sentinel-1 observations were matched to in situ soil moisture by station and date (same day). SMOSMANIA in situ data were obtained from the International Soil Moisture Network [10], retaining only observations flagged as good ('G') at the shallowest layer (5 cm). Station-mean soil moisture and the error magnitudes derived from it (Table 1; Figure 2 ) are uniformly expressed in $m^3\ m^{-3}$: OzNet and SMOSMANIA values, which are natively in %vol, are divided by 100, whereas REMEDHUS and Little Washita are natively in $m^3\ m^{-3}$ and left unchanged. Observations outside physically plausible ranges of backscatter, in situ soil moisture, and NDVI were excluded (thresholds in Supplementary Section S1).

#### 2.2.2. WCM Decoupling

The decoupled feature VV_soil is computed using the standard Attema–Ulaby formulation, given by

$$T^2 = \exp(-2B \cdot NDVI/\cos\theta) \quad (1)$$

$$\sigma_veg = A \cdot LAI \cdot \cos\theta \cdot (1 - T^2) \quad (2)$$

$$\sigma_soil = (\sigma_total - \sigma_veg)/T^2 \quad (3)$$

where T denotes the two-way vegetation transmissivity; θ is the local incidence angle; A and B are empirical vegetation-scattering and -attenuation coefficients, calibrated as described in Section 2.2.3; NDVI is the Normalized Difference Vegetation Index; LAI is the leaf area index, approximated as LAI_proxy = NDVI × 3 (clipped to [0, 6]); and σ_veg, σ_soil, and σ_total denote the vegetation, soil, and total backscattering coefficients, respectively. The same decoupling (Equations (1)–(3), with LAI_proxy = NDVI × 3) was applied to all four networks.

The vegetation parameters differ among the networks. OzNet and SMOSMANIA adopt A_VV = 0.10 and A_VH = 0.05, and Little Washita uses the same values after independent verification (Section 4.5). REMEDHUS uses region-specific literature parameters, A_VV = 0.0046 and A_VH = 0.0017 (Baghdadi et al. [12], calibrated for winter-cereal

canopies), because a shared A_VV = 0.10 was found not to converge for this network. The incidence angle θ is the actual per-observation value at REMEDHUS, OzNet, and SMOSMANIA (approximately 30–46°); at Little Washita, the fixed representative angle of 39° provided by the established station-level dataset is used. The implications of these parameterization differences for the comparison among networks are discussed in Section 4.5.

#### 2.2.3. Calibration of the B Parameter

The B parameter was determined by grid search. Within a plausible range (step size 0.025–0.05, varying by region), the Pearson correlation between σ_soil (VV) and SM_insitu was computed at both the pooled (all-station) and per-station scales, and the region-wide optimal B was used in the main analysis. At REMEDHUS and Little Washita, the optimum was confirmed with an expanded grid (0.05–20) to ensure that it is a genuine interior peak rather than a boundary artifact, giving B = 1.900 (REMEDHUS) and 0.900 (Little Washita) (Table 1; search history in Supplementary Section S2). The decoupled VV_soil was clipped at the 1st–99th percentiles to control for extreme values.

Per-station optimal B shows substantial dispersion within each region, suggesting that the spatial heterogeneity of B is itself a potential error source (Section 4.5). As a robustness check, B calibration was repeated within each LOSO fold using only training-station data (leakage-free nesting) instead of the pooled all-station calibration; the resulting M3 $R^2$ values were nearly identical (max Δ = 0.0017, REMEDHUS), indicating that B calibration is unlikely to introduce a material leakage artifact in this study.

At OzNet and SMOSMANIA, the grid search did not identify a clear optimum. At OzNet, the correlation was highest at the lowest grid value ($r = +0.10$ at $B = 0.05$) and negative for every $B \geq 0.15$, so there was no interior peak. At SMOSMANIA, it rose from negative values at small B ($r = -0.15$ at $B = 0.3$) to a weak, nearly flat plateau ($r \approx +0.10$ for $B \gtrsim 5$, with a maximum of +0.104 at B = 9.5 on the expanded 0.05–20 grid; the maximum within the standard grid, +0.06, lay on its upper boundary), with no distinct peak (Supplementary Section S2). Literature-default values were therefore used (B = 0.300 for OzNet and B = 0.243 for SMOSMANIA; marked 'uncalibrated' in Table 1). At both networks, the decoupled VV/VH correlates less with SM than the raw bands (SMOSMANIA: ΔrVV = −0.272, ΔrVH = −0.199; OzNet: ΔrVV = −0.361, ΔrVH = −0.374); see Section 4.5. A sensitivity test of the vegetation-scattering term is given in Supplementary Section S3.

The analyses in this study differ in feature set, validation scheme, and random forest (RF) configuration. Table 2 summarizes, for each experiment, the feature set, validation approach, RF configuration, and purpose, so that the following sections can be read against it.

**Table 2.** Overview of the analyses: feature set, validation scheme, RF configuration, and purpose.

| Experiment | Section | Feature Set | Validation | RF Configuration | Purpose |
|---|---|---|---|---|---|
| Random-split bootstrap | Section 3.1 | Canonical M1–M3 (Section 2.3) | Random 80/20 split; bootstrap resampling (50/20/50/12 resamples at REMEDHUS/Little Washita/OzNet/SMOSMANIA); Wilcoxon test on paired errors (M2 vs. M3) | Bootstrap: $n = 80$, depth = 8; Wilcoxon: $n = 200$, depth = 10 (leaf = 5) | Test whether the WCM feature helps under random validation |
| Canonical LOSO | Section 3.2 | Canonical M1–M3 (Section 2.3) | LOSO; compared with climatology/persistence baselines | Canonical | Evaluate cross-site generalization |
| Anomaly modeling | Section 3.3 | Canonical M3 | LOSO; absolute vs. leaky vs. leakage-free anomaly | Canonical | Quantify station-mean leakage |
| Error decomposition | Section 3.4 | Canonical M1–M3 | LOSO | Canonical | Separate station-mean bias from temporal RMSE |

**Table 2.** *Cont*.

| Experiment | Section | Feature Set | Validation | RF Configuration | Purpose |
|---|---|---|---|---|---|
| Fixed-n subnetwork | Section 3.5 | Canonical M3 | LOSO within random 6-station subnetworks (300 per region; 924 at OzNet) | Lightweight and canonical | Test the CV–$R^2$ relation and hyperparameter sensitivity |
| OzNet mechanism | Section 3.6 | Canonical M3 | LOSO over 924 OzNet subnetworks | Lightweight | Mediation-path analysis |
| Domain correction | Section 3.7 | Canonical M3; M4 (Section 2.3) | LOSO, scored on the later 90% of the held-out station | Canonical | Test whether the distribution shift is correctable |

Canonical feature set: M1 = VV_dB, VH_dB, VV/VH; M2 = M1 + NDVI, VV × NDVI, VH × NDVI; M3 = M2 + VV_soil (Section 2.3). Canonical RF: $n$ = 100, depth = 10, leaf = 5, seed = 42. Lightweight RF: $n$ = 40, depth = 6. LOSO: leave-one-station-out. Random-split RF: $n$ = 80, depth = 8, leaf = 5 for the bootstrap $R^2$; $n$ = 200, depth = 10, leaf = 5 for the Wilcoxon test. RMSE: root-mean-square error.

### *2.3. Model, Features, and Validation Framework*

**Feature Sets**

The WCM-decoupled feature VV_soil is the dB-converted soil backscatter, VV_soil = 10·lg($\sigma$_soil), where $\sigma$_soil is obtained from Equations (1)–(3) in the linear domain ($\sigma$_total = $10^{(VV/10)}$). The canonical feature set (M3; 7 dimensions) comprises VV_dB, VH_dB, VV/VH, NDVI, VV × NDVI, VH × NDVI, and VV_soil; M1 uses the first three dimensions and M2 the first six. This canonical M1/M2/M3 hierarchy is used throughout Sections 3.1–3.7. M4 (see "Station-Mean Feature Model (M4)" in this section) is used only in Section 3.7.

The same nested feature sets (M1 ⊂ M2 ⊂ M3; Table 3) are used in both the random-split analysis (Section 3.1) and the LOSO analyses (Sections 3.2–3.7).

**Table 3.** Input variables of the M1, M2, and M3 models used in both the random-split (Section 3.1) and LOSO (Sections 3.2–3.7) analyses.

| Model | Random-Split Validation (Section 3.1) | LOSO Validation (Sections 3.2–3.7) |
|---|---|---|
| M1 | VV_dB, VH_dB, VV/VH (3 variables) | VV_dB, VH_dB, VV/VH (3 variables) |
| M2 | M1 + NDVI, VV × NDVI, VH × NDVI (6 variables) | M1 + NDVI, VV × NDVI, VH × NDVI (6 variables) |
| M3 | M2 + VV_soil (7 variables) | M2 + VV_soil (7 variables) |

Note: VV_dB and VH_dB are Sentinel-1 backscatter coefficients (dB); VV/VH is the cross-polarization ratio; NDVI is the Sentinel-2 normalized difference vegetation index; VV_soil is the WCM-decoupled soil backscatter (dB); × denotes an interaction term. The feature sets are strictly nested (M1 ⊂ M2 ⊂ M3) and identical across the four networks and the two validation schemes.

Figure 3 summarizes the end-to-end pipeline, from raw Sentinel-1/2 inputs through WCM decoupling and the three canonical feature sets to the random forest model and the two validation schemes (random-split and LOSO; see Validation Framework below).

**Random Forest Configuration**

All LOSO figures are computed under the same canonical parameters: RF ($n$ = 100, depth = 10, leaf = 5, seed = 42); an upper cap of 5000 training samples for Little Washita and SMOSMANIA; data filtering applies dropna only to VV_dB/NDVI/SM, ensuring full comparability.

The applicability and hyperparameter sensitivity of random forest have been reviewed in detail [55,56]; support vector machines [57] and standard statistical-learning texts [58] provide broader context for method selection. Interpretability tools such as SHapley Additive exPlanations (SHAP) [59,60] and the theory-guided data science paradigm [61] could

be used in future work to dissect the marginal contribution of the WCM features, but were not adopted here for methodological consistency.

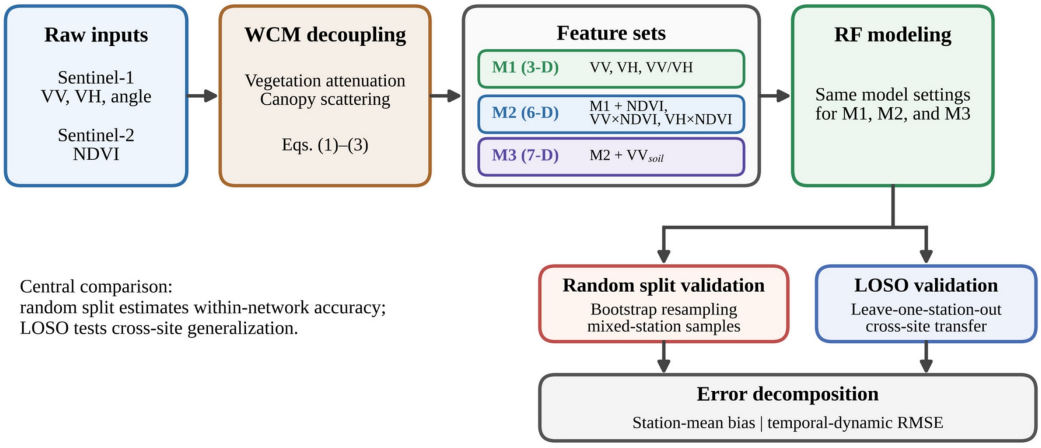



**Figure 3.** Overview of the WCM-based feature construction, random forest modeling, and dual validation (random-split vs. LOSO) pipeline, culminating in the station-mean-bias/temporal-RMSE error decomposition described in Section 2.4.

**Validation Framework (Canonical Pipeline)**

Three components of the canonical validation framework are described below: LOSO validation, leakage-free anomaly modeling, and the climatology baseline.

(1) **LOSO Validation**

In each fold, all observations from one station are completely removed from the training set; the model is trained on the remaining stations and used to predict the held-out station, and this is repeated across all stations before aggregating the results. LOSO rigorously evaluates a model's ability to extrapolate to stations unseen during training—i.e., cross-site generalization—which is an appropriate evaluation criterion for real-world, network-scale deployment, and more faithfully reveals a model's spatial extrapolation error than random splitting.

(2) **Leakage-Free Anomaly Modeling**

Anomaly modeling (SM_anomaly) removes each station's climatological baseline and models the remaining temporal fluctuations; it is commonly used to improve SM retrieval accuracy. However, if the test station's data are included when the station means are computed, information leaks from the test set and the apparent performance is inflated. This study therefore computes a leaky and a leakage-free version to quantify and correct this effect:

With leakage: station means computed from all data (including the test station) → leakage.

Leakage-free (corrected): means computed from training stations only → the test station uses the regional mean from the training set instead → pred_abs = pred_anomaly + regional_mean. The share of the apparent gain attributable to leakage is defined relative to absolute-value modeling (M3), because this directly answers how much of the apparent gain, compared with not performing anomaly modeling at all, is spurious.

(3) **Station-Mean Feature Model (M4)**

To test whether the station-baseline shift can be learned in feature space rather than corrected after prediction, M4 adds one feature to M3, the station's mean in situ SM (8 dimensions in total). Training stations use their own mean. For the held-out station, two versions were evaluated: M4-pool receives the mean of the training stations' means (the leakage-free convention above, so no information about the held-out station is used), and

M4-warm receives the mean of its earliest 10% of observations, i.e., the same information used by the target-mean correction of Section 3.7. Both versions use the canonical RF configuration and the LOSO protocol, and are scored on the later 90% of each held-out station's observations, as in Section 3.7.

(4) **Climatology Baseline**

Monthly mean SM is computed at the training stations and applied to the same month at the test station. This baseline uses no cross-station modeling and relies only on monthly climatological regularities, so it serves as a strict, leakage-free reference. If an RF model cannot outperform it, the model has learned no transferable signal under LOSO.

### *2.4. Error Decomposition Framework*

LOSO error is decomposed into two components. Station-mean bias (mean_abs_bias) reflects whether the model has learned each station's absolute SM baseline level. Temporal-dynamic RMSE, the RMSE of anomalies after removing the station mean, reflects whether the model has learned the temporal dynamics of SM. The decomposition separates two distinct prediction tasks, 'static baseline localization' and 'dynamic process characterization', so that the contribution of the WCM-decoupled feature can be assessed for each. It follows the same logic as the triple-collocation error decomposition widely used in remote-sensing product validation [62–64]: total error is partitioned into attributable components so that the marginal effect of model improvements can be located more precisely.

### *2.5. Fixed-n Subnetwork Analysis and RF Hyperparameter Sensitivity*

With $n$ fixed at 6 stations, 300 random combinations were drawn per region (924 exhaustive combinations for OzNet), and the subset SM mean CV and LOSO $R^2$ were computed for each. Two RF configurations were used: a lightweight RF ($n = 40$, depth = 6, for large-scale sampling) and a canonical RF ($n = 100$, depth = 10, 40 re-checks per region); robustness of the conclusions was judged by directional agreement between the two sets of results.

## 3. Results

### *3.1. Random Validation (Bootstrap Resampling, RF)*

Under random-split bootstrap validation, mean $R^2$ for the M1/M2/M3 models is low to moderate across the four regions (REMEDHUS 0.03–0.10, Little Washita 0.11–0.13, OzNet Yanco 0.05–0.12, SMOSMANIA 0.09–0.28; Table 4; Figure 4a), and the improvement of the WCM-decoupled feature (M3) over the base SAR model (M2) is not statistically significant in any of the four regions (Wilcoxon signed-rank test, $p = 0.168$–$0.982$; Table 4). The mean differences are also small in magnitude ($\Delta R^2$ = M3 − M2: +0.003 at REMEDHUS, +0.000 at Little Washita, −0.001 at OzNet Yanco, and +0.001 at SMOSMANIA; Table 4), at most 0.003 in absolute value and well within the overlapping 95% confidence intervals of M2 and M3, and much smaller than the M1-to-M2 gain at all four networks (+0.073, +0.022, +0.063, and +0.187 at REMEDHUS, Little Washita, OzNet Yanco, and SMOSMANIA, respectively).

The non-significant $p$-values therefore correspond to differences that are also small in magnitude, rather than to large but imprecisely estimated differences. The bootstrap resampling method [65] and the Wilcoxon signed-rank test [66] used in this section are both classical nonparametric statistical methods, appropriate when the sample distribution is unknown or non-normal and not reliant on asymptotic normality assumptions.

**Table 4.** Random-split bootstrap validation results (RF Algorithm, M2 vs. M3 Wilcoxon Signed-Rank Test).

| Region | Model | Mean $R^2$ | $R^2$ 95% CI | M2 vs. M3 *p*-Value | $\Delta R^2$ (M3 − M2) |
|---|---|---|---|---|---|
| REMEDHUS | M1/M2/M3 | 0.028/0.101/0.104 | [−0.009,0.062]/[0.059,0.144]/[0.068,0.147] | $p = 0.168$ ns | +0.003 |
| Little Washita | M1/M2/M3 | 0.111/0.133/0.133 | [0.092,0.134]/[0.116,0.159]/[0.115,0.160] | $p = 0.750$ ns | +0.000 |
| OzNet Yanco | M1/M2/M3 | 0.055/0.118/0.117 | [−0.012,0.104]/[0.070,0.157]/[0.068,0.156] | $p = 0.911$ ns | −0.001 |
| SMOSMANIA | M1/M2/M3 | 0.093/0.280/0.281 | [0.080,0.102]/[0.264,0.296]/[0.266,0.297] | $p = 0.982$ ns | +0.001 |

Note: $R^2$ and *p*-values are from the Random Forest (RF) algorithm; the Wilcoxon signed-rank test is based on paired comparison of per-sample absolute errors. Number of bootstrap resamples: 50 each for REMEDHUS and OzNet, 20 for Little Washita, and 12 for SMOSMANIA (the number of resamples decreases roughly with increasing sample size, to control computation time; Table 4 shows that confidence intervals remain stable across the four regions under this setting, and this does not affect the direction of the significance conclusions; minimum $p = 0.168$). $\Delta R^2$ = mean $R^2$ of M3 minus mean $R^2$ of M2 (computed from the tabulated means); positive values indicate that M3 is better. ns, not significant ($p \geq 0.05$).

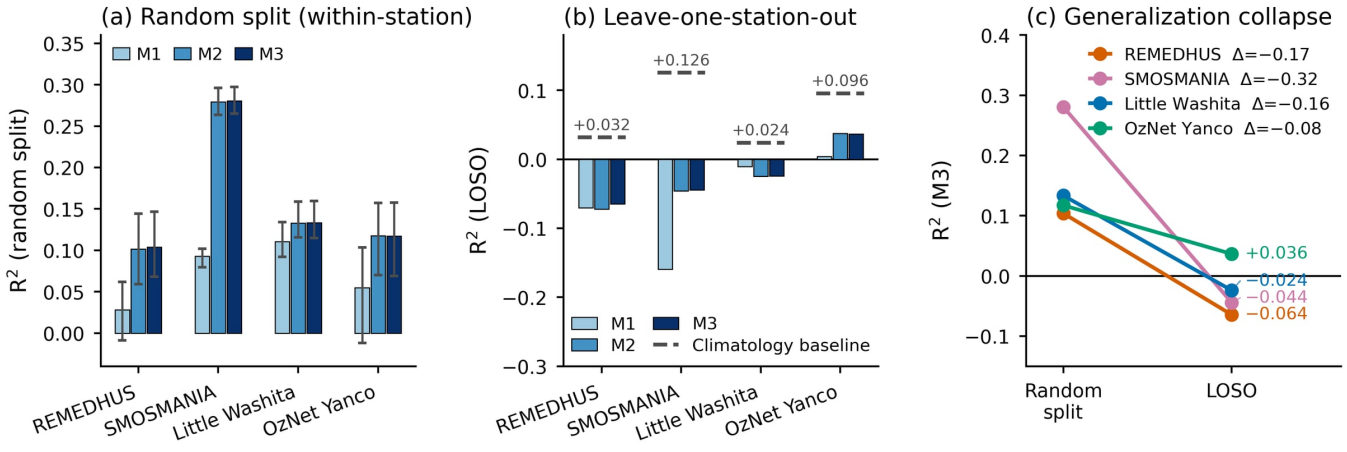



**Figure 4.** Comparison of random-split and LOSO validation. (**a**) Mean $R^2$ and 95% bootstrap confidence intervals for M1/M2/M3 across the four regions under within-station random splitting; (**b**) $R^2$ under LOSO validation, with the gray dashed line indicating the Climatology baseline; (**c**) performance drop for the M3 model from random split to LOSO. All RF models fall below the Climatology baseline under LOSO. SMOSMANIA exhibits both the highest random-split $R^2$ among the four regions (M3 ≈ 0.28) and, despite a still-negative LOSO $R^2$ (M3 = −0.04), the largest performance drop from random split to LOSO among the four regions ($\Delta = -0.32$).

### *3.2. LOSO Validation (Canonical Pipeline)*

Under LOSO cross-site validation, $R^2$ collapses across all four regions to near zero or negative, falling below the monthly climatology baseline in every case (Table 5; Figure 4b,c), suggesting that what the models learn during training is largely 'station memory' with limited cross-site transferability; only OzNet's M3 records a weak positive value (+0.036), which still falls short of its own climatology baseline (+0.096). SMOSMANIA additionally exhibits the largest random-split-to-LOSO performance drop among the four regions ($\Delta = -0.32$, from M3 ≈ 0.28 under random splitting to M3 = −0.04 under LOSO; Figure 4), providing additional supporting evidence that random splitting overestimates cross-site generalization ability. SMOSMANIA's M1 is also the most negative of the four regions (−0.159; Table 5). At SMOSMANIA, where B could not be calibrated to a clear optimum (Section 2.2.3), M3 shows almost no improvement over M2 (−0.044 vs. −0.046; Table 5). Across the four networks, the LOSO $\Delta R^2$ (M3 − M2) is +0.008 (REMEDHUS), +0.001 (Little Washita), −0.001 (OzNet Yanco), and +0.002 (SMOSMANIA), all within ±0.01 and far smaller than the gap between M3 and the climatology baseline at each network. The

Persistence baseline (within-station temporal), by contrast, is far higher than any cross-site method (Table 5), indicating strong temporal autocorrelation that does not transfer across stations [67].

**Table 5.** LOSO validation results compared with the climatology/persistence baselines.

| Region | M1 | M2 | M3 | Climatology | Persistence | Landscape CV |
|---|---|---|---|---|---|---|
| REMEDHUS | −0.070 | −0.072 | −0.064 | +0.032 | +0.827 | 0.673 |
| Little Washita | −0.010 | −0.025 | −0.024 | +0.024 | +0.721 | 0.364 |
| OzNet Yanco | +0.004 | +0.037 | +0.036 | +0.096 | +0.321 | 0.184 |
| SMOSMANIA | −0.159 | −0.046 | −0.044 | +0.126 | +0.898 | 0.380 |

Note: (1) all RF models have LOSO $R^2$ below climatology; (2) Persistence (within-station temporal) is far higher than any cross-site method; (3) OzNet's M3 $R^2 = +0.036 > 0$ represents only weak positive generalization, still below its climatology baseline (+0.096).

### *3.3. Anomaly Modeling: Leakage Correction*

After re-modeling with station-level anomalies (temporal fluctuations after removing each station's climatological mean), the $R^2$ under leaky anomaly modeling differs dramatically from absolute-value modeling across the four regions (REMEDHUS +0.703, Little Washita +0.669, OzNet Yanco +0.306, SMOSMANIA +0.630; Table 6; Figure 5a); however, after removing station-mean leakage, only OzNet and SMOSMANIA retain a meaningful leakage-free improvement (leakage-free $R^2$ of +0.226 and +0.171, versus +0.036 and −0.044 for absolute-value modeling), while the gains at REMEDHUS (+0.044 versus −0.064) and Little Washita (+0.033 versus −0.024) are almost entirely attributable to data leakage. Quantitatively, station-mean leakage accounts for 86% and 92% of the apparent gain over absolute-value modeling at REMEDHUS (leakage-free $R^2$ +0.044) and Little Washita (+0.033), compared with 30% at OzNet and 68% at SMOSMANIA (Table 6; Figure 5b). The leakage share is not monotonically related to landscape CV: SMOSMANIA (CV = 0.380) and Little Washita (CV = 0.364) have similar CV but leakage shares of 68% and 92%, and the leakage-free $R^2$ at SMOSMANIA (+0.171) is close to that at OzNet (+0.226) and far higher than at Little Washita (+0.033).

**Table 6.** Leakage correction for station-mean bias in anomaly modeling.

| Region | M3 Absolute | Leaky Anom + DOY | Leakage-Free Anom + DOY | Share of Apparent Gain due to Leakage * |
|---|---|---|---|---|
| REMEDHUS | −0.064 | +0.703 | +0.044 | 86% |
| Little Washita | −0.024 | +0.669 | +0.033 | 92% |
| OzNet Yanco | +0.036 | +0.306 | +0.226 | 30% |
| SMOSMANIA | −0.044 | +0.630 | +0.171 | 68% |

Note: * Share = 1 − ($R^2$_leakage-free − $R^2$_M3)/($R^2$_leaky − $R^2$_M3), i.e., the fraction of the apparent gain over absolute-value modeling (M3) that is attributable to leakage. Under the alternative ratio-based convention, 1 − $R^2$_leakage-free/$R^2$_leaky, the shares would be 94%, 95%, 26%, and 73% for REMEDHUS, Little Washita, OzNet, and SMOSMANIA, respectively. Anom + DOY, anomaly modeling with day-of-year (DOY) terms added as predictors.

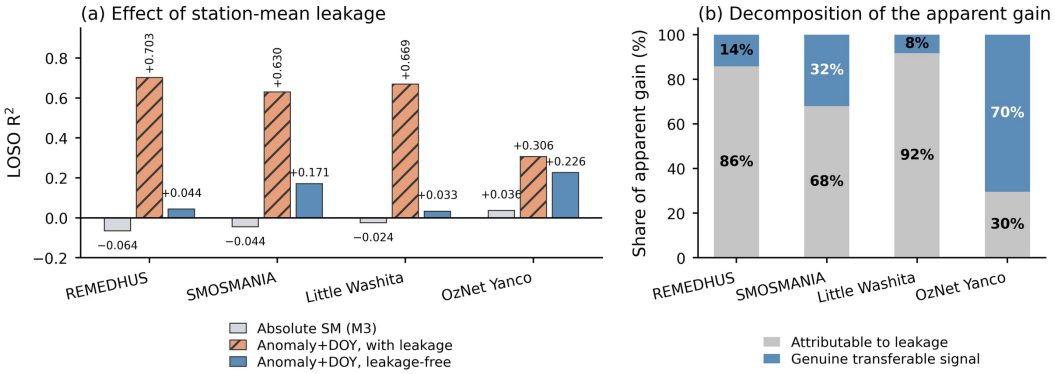



**Figure 5.** Station-mean leakage in anomaly modeling and its correction. (**a**) LOSO $R^2$ under three conventions: absolute-value modeling (M3), leaky anomaly modeling with day-of-year (DOY) terms (hatched), and leakage-corrected anomaly + DOY modeling; (**b**) decomposition of the apparent gain, with gray indicating the leakage contribution.

### *3.4. Error Decomposition*

After decomposing LOSO error into station-mean bias and temporal-dynamic RMSE, M1/M2/M3 are highly similar on both components across all four regions (Table 7; Figure 6), suggesting that the WCM-decoupled feature does not compress the systematic baseline bias between stations or systematically improve the model's ability to capture temporal dynamics. Taken together, the WCM-decoupled feature (M3) yields no significant improvement over M2 along any of the four dimensions examined: random-split validation ($p \geq 0.168$), LOSO $R^2$, station-mean bias, and temporal-dynamic RMSE (Tables 4, 5 and 7).

**Table 7.** Two -component decomposition of LOSO error.

| Region | Model | Station-Mean Bias (MAE, $m^3$ $m^{-3}$) | Temporal-Dynamic RMSE ($m^3$ $m^{-3}$) | Overall LOSO $R^2$ | No. of Stations |
|---|---|---|---|---|---|
| REMEDHUS | M1/M2/M3 | 0.0651/0.0651/0.0648 | 0.0558/0.0570/0.0568 | −0.070/−0.072/−0.064 | 19 |
| Little Washita | M1/M2/M3 | 0.0605/0.0616/0.0616 | 0.0624/0.0628/0.0627 | −0.010/−0.025/−0.024 | 21 |
| OzNet Yanco | M1/M2/M3 | 0.0231/0.0235/0.0233 | 0.0715/0.0704/0.0705 | +0.004/+0.037/+0.036 | 12 |
| SMOSMANIA | M1/M2/M3 | 0.0608/0.0506/0.0506 | 0.0776/0.0797/0.0796 | −0.159/−0.046/−0.044 | 21 |

Note: MAE, mean absolute error between each station's predicted and observed mean, averaged across stations.

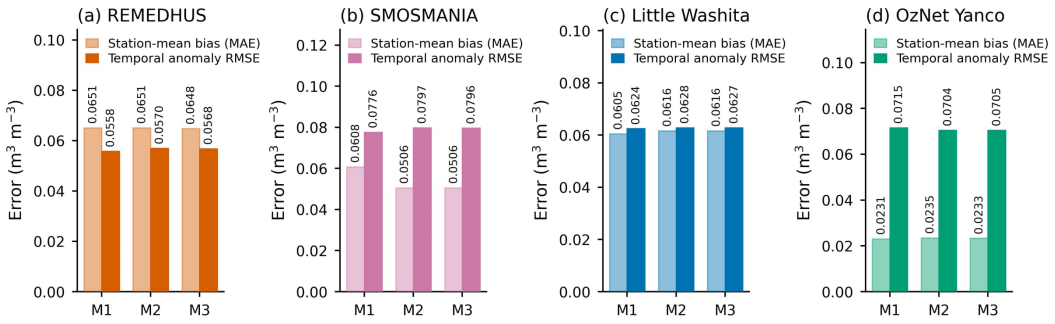



**Figure 6.** Two-component decomposition of LOSO error. Light bars show station-mean bias (the absolute difference between each station's predicted and observed mean, averaged across stations); dark bars show temporal-dynamic RMSE (the RMSE of anomalies after removing each station's own mean, averaged across stations). M1, M2, and M3 are highly similar on both components across all four regions. All panels are expressed in $m^3$ $m^{-3}$; the vertical axes differ among panels. (**a**) REMEDHUS; (**b**) SMOSMANIA; (**c**) Little Washita; (**d**) OzNet Yanco.

### *3.5. Fixed-n Subnetwork Analysis: Comparison of Two RF Configurations*

LOSO $R^2$ decreases with landscape heterogeneity across the four networks (Table 5; Figure 2b): REMEDHUS (CV = 0.673, $R^2 = -0.064$) is the lowest, followed by SMOSMANIA (CV = 0.380, $R^2 = -0.044$), Little Washita (CV = 0.364, $R^2 = -0.024$), and OzNet (CV = 0.184, $R^2 = +0.036$), so that the four networks form a monotonic CV–$R^2$ ordering. SMOSMANIA and Little Washita, which have similar CV levels, are adjacent in this ordering.

The fixed-n ($n = 6$) subnetwork resampling analysis shows that the strength and direction of the relationship between subnetwork SM CV and LOSO $R^2$ vary by region (Table 8; Figure 7a–d): only REMEDHUS maintains a robust negative correlation under both the lightweight and canonical RF configurations (canonical RF: r = −0.453, descriptive $p = 0.003$; Table 8); Little Washita, SMOSMANIA, and OzNet each show a hyperparameter-sensitive pattern—a significant correlation under lightweight RF that vanishes, or for Little Washita reverses sign, and loses significance under canonical RF. Figure 7e compares the Spearman r values under the two RF configurations directly.

**Table 8.** Fixed -n ($n = 6$) subnetwork analysis: hyperparameter sensitivity of lightweight vs. canonical RF.

| Region | N | Lightweight RF Spearman r | *p*-Value | Canonical RF Spearman r | *p*-Value (N = 40) | Direction Stable |
|---|---|---|---|---|---|---|
| REMEDHUS (CV = 0.673) | 300 | −0.286 | <0.001 *** | −0.453 | 0.003 ** | Yes—robust |
| Little Washita (CV = 0.364) | 300 | −0.132 | 0.022 * | (+0.121) | 0.456 ns | No—not replicated |
| OzNet Yanco (CV = 0.184, exhaustive) | 924 | +0.137 | <0.001 *** | −0.018 | 0.912 ns | No—not replicated |
| SMOSMANIA (CV = 0.380) | 300 | −0.262 | <0.001 *** | −0.069 | 0.673 ns | No—not replicated |

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ns, not significant. Because the 300 subnetworks (924 at OzNet) are drawn from the same set of stations and therefore overlap, they are not statistically independent; the *p*-values are descriptive indicators of consistency, not formal significance tests.

### *3.6. Supplementary Mechanistic Analysis of OzNet (Under Lightweight RF)*

Given the potential mechanistic value of the positive correlation observed at OzNet under lightweight RF, a mediation-path analysis was conducted across the 924 combinations: does a higher-CV subnetwork actually provide broader training coverage? Results show that CV is highly correlated with the training SM mean range (Spearman r = +0.973, $p < 0.001$) (Figure 8a), and the training SM mean range is positively correlated with LOSO $R^2$ (r = +0.121, $p < 0.001$) (Figure 8b). However, the VV mean range (r = −0.525) and NDVI mean CV (r = −0.266) are negatively correlated with LOSO $R^2$, indicating that VV/NDVI feature-domain coverage is not a positive mediator. It should be noted that the metric used here, the range of training-station SM means, is highly correlated with CV and weakly positively correlated with LOSO $R^2$ (Figure 8c), providing a partial mediation pathway; other SM-range-type metrics do not consistently support this pathway. Combined with the fact that the positive correlation vanishes under canonical RF, this correlation is sensitive to RF hyperparameters and cannot be regarded as a stable landscape mechanism.

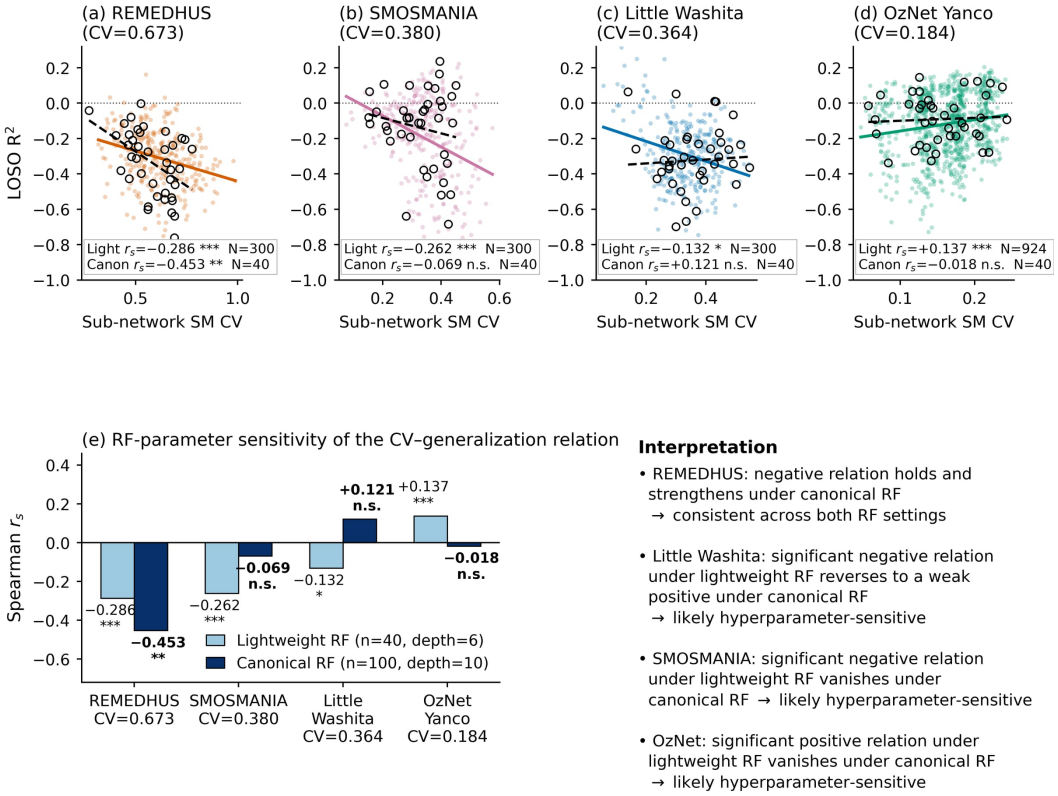



**Figure 7.** Fixed-n ($n$ = 6) subnetwork analysis and its RF-hyperparameter sensitivity. (**a**–**d**) Relationship between subnetwork SM CV and LOSO $R^2$ across the four regions: solid dots and solid lines show the large-sample results under lightweight RF ($n$ = 40, depth = 6); open circles and black dashed lines show the 40-repeat re-check results under canonical RF ($n$ = 100, depth = 10); (**e**) comparison of Spearman correlation coefficients under the two RF configurations. REMEDHUS's negative correlation is directionally consistent and significant under both configurations; Little Washita, SMOSMANIA, and OzNet each show a significant correlation under lightweight RF (negative for Little Washita and SMOSMANIA, positive for OzNet) that vanishes or reverses sign under canonical RF. In the boxes in (**a**–**d**), Light and Canon denote lightweight and canonical RF, r_s is the Spearman rank correlation coefficient, and N is the number of subnetworks; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s., not significant.

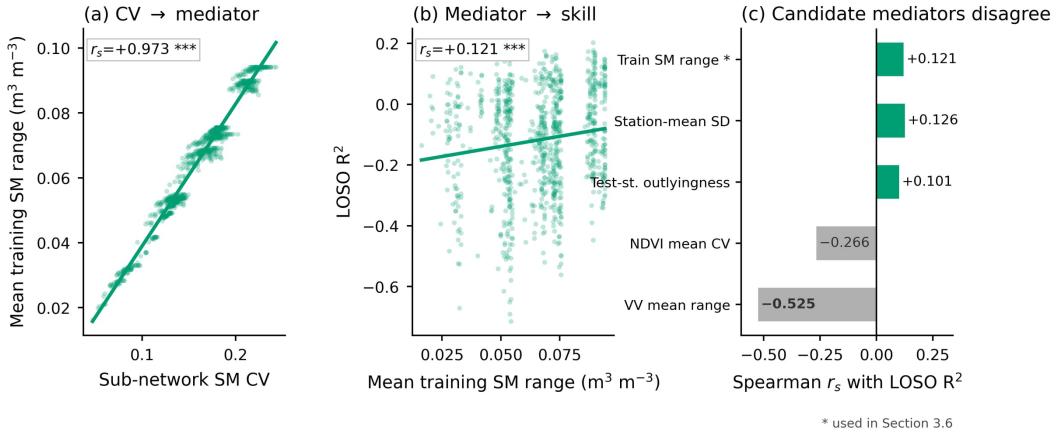



**Figure 8.** Mediation-path analysis of the OzNet subnetworks (N = 924, under lightweight RF). (**a**) Subnetwork CV is highly positively correlated with the training-station SM mean range; (**b**) this metric is weakly positively correlated with LOSO $R^2$; (**c**) comparison of Spearman correlations between five candidate mediator metrics and LOSO $R^2$, with the starred metric being the one adopted in the main text (Section 3.6 ). r_s, Spearman rank correlation coefficient; *** $p < 0.001$.

### *3.7. Domain-Correction Contrast: Supporting Evidence for Distributional Shift*

Sections 3.2–3.6 indicate that WCM-ML's cross-site failure is consistent with distributional shift rather than vegetation-correction error, but a purely correlational diagnosis leaves open whether this shift is actually correctable. To examine this experimentally, five settings were compared under the same canonical M3 features and RF configuration, differing only in how (or whether) the held-out station's distribution is aligned to the training domain: no correction; target-station-mean correction (shifting predictions using only the mean of the station's earliest 10% of observations, simulating a brief calibration period at deployment; hereafter "target-mean correction"); feature standardization (CORrelation ALignment (CORAL)-style z-scoring of train and test data separately); few-shot fine-tuning (folding the same earliest 10% of samples into training); and the two combined. Evaluation is always restricted to the remaining (later) 90% of the held-out station's observations, identical across all five settings, so no setting sees any of its own evaluation-period labels.

Target-station-mean correction alone lifts REMEDHUS, Little Washita, and SMOSMANIA from negative LOSO $R^2$ to +0.291, +0.332, and +0.159 respectively, in all three cases exceeding that region's own climatology baseline (Table 9; Figure 9a), while degrading OzNet's near-zero LOSO performance (from +0.041 to −0.132). Feature standardization and few-shot fine-tuning, by contrast, provide little or no improvement at any site, and few-shot fine-tuning is diluted by the remaining 11–20 training stations. A large, one-directional distributional correction succeeding where feature-level or sample-level adjustments fail is a pattern consistent with a cross-site failure associated with a shift in each station's absolute SM baseline rather than with any deficiency in the SAR–vegetation feature relationship itself, consistent with the error-decomposition result in Section 3.4. The gain from target-mean correction is also directionally consistent with landscape heterogeneity (Figure 9b): positive at all three heterogeneous-to-intermediate networks (CV $\geq$ 0.36) and negative at the one homogeneous network (OzNet, CV = 0.18), where station baselines are already close enough that an estimated correction merely adds noise (Pearson r = 0.771 across the four sites; with only four networks this is reported descriptively, not as a significance test). A station-level version of the same test, pooling all 73 individual held-out stations across the four networks and replacing landscape CV with each station's own baseline outlyingness relative to its training pool (Figure 9c), is directionally consistent with this, based on a hypothesis test rather than a four-point description (Spearman $\rho = 0.342$, $p = 0.003$, $n = 73$). Because the correction uses only the earliest samples from each held-out station, mimicking a short in situ calibration period at a newly deployed site, this suggests a potentially deployable, low-cost remedy, although it was tested retrospectively and only on the four networks examined.

To quantify how long the calibration window must be, the target-mean correction was repeated with the warm-up window set to the earliest 2, 5, 10, 20, 30, and 50% of each held-out station's observations (Figure 10). Because the 50% window leaves only the later half of each record unused, all windows were scored on the same last 50% of the record, so that $R^2$ is comparable across windows; the 10% values therefore differ from Table 9, which is scored on the later 90% (Table 9 is reproduced exactly when the original protocol is applied). At Little Washita and SMOSMANIA, the corrected $R^2$ exceeded the uncorrected value at every window tested, including 2–5% (Little Washita +0.19 to +0.20 vs. −0.01; SMOSMANIA +0.14 to +0.17 vs. −0.09), and rose from roughly +0.14–0.20 at the shortest windows to +0.27–0.34 at the 50% window. At REMEDHUS, windows of 5% or shorter were no better than no correction (−0.20 and −0.15 vs. −0.09), whereas a 10% window already gave +0.22 and 20% or longer gave +0.36 to +0.41. At OzNet, correction was worse than no correction at every window (−0.51 at 2%, improving to −0.10 at 50% against −0.02 uncorrected), so a longer window reduced but never reversed the harm.

Taken together, the gain increases with window length and is not an artifact of the particular 10% choice, which is consistent with the gain arising from estimating each station's mean offset more accurately; the dependence is not strictly monotonic (e.g., SMOSMANIA at 5% vs. 10%), and with 12–21 stations per network, small fluctuations should not be over-interpreted. As a practical rule, a calibration window of at least about 10% of the station's record was needed for the correction to help at all three heterogeneous networks.

**Table 9.** Domain-correction contrast across four validation networks (M3, Canonical RF, LOSO R2). Evaluation is restricted to the later 90% of each held-out station's observations, so the "No correction" values differ slightly from Table 5.

| Region | No Correction | Target-Mean Correction | Feature Std. (CORAL) | Few-Shot Fine-tune | TgtMean + FeatStd | Climatology Baseline |
|---|---|---|---|---|---|---|
| REMEDHUS (CV = 0.673) | −0.062 | +0.291 | −0.129 | −0.043 | +0.218 | +0.032 |
| Little Washita (CV = 0.364) | −0.023 | +0.332 | −0.020 | +0.003 | +0.325 | +0.024 |
| OzNet Yanco (CV = 0.184) | +0.041 | −0.132 | +0.047 | +0.048 | −0.140 | +0.096 |
| SMOSMANIA (CV = 0.380) | −0.055 | +0.159 | −0.156 | +0.028 | +0.042 | +0.126 |

Note: TgtMean + FeatStd, target-station-mean correction combined with feature standardization.

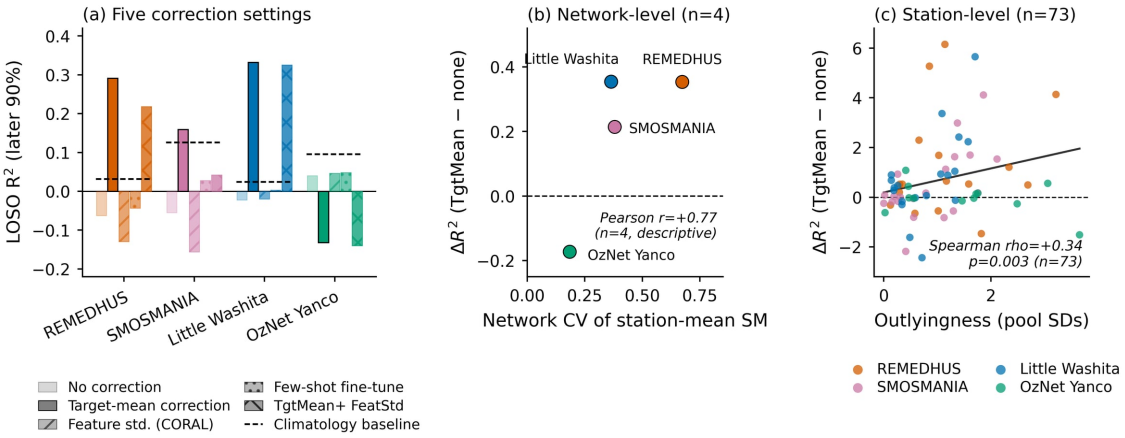



**Figure 9.** Domain-correction contrast across the four validation networks. (**a**) LOSO $R^2$ under five correction settings (M3, canonical RF), with each region's Climatology baseline shown as a dashed line and target-station-mean correction shown as solid bars. (**b**) Gain from target-mean correction ($\Delta R^2$ relative to no correction) plotted against each network's landscape CV. (**c**) The same relationship re-expressed at station level ($n = 73$, pooled across all four networks): each held-out station's baseline outlyingness relative to its training pool (warmup-period mean only, no evaluation-label leakage) against its own target-mean-correction gain, with a Spearman rank test replacing the network-level descriptive correlation in (**b**). Outlyingness is the absolute difference between a held-out station's warm-up mean and the training-pool mean, expressed in standard deviations of the training-pool station means. TgtMean + FeatStd denotes target-station-mean correction combined with feature standardization.

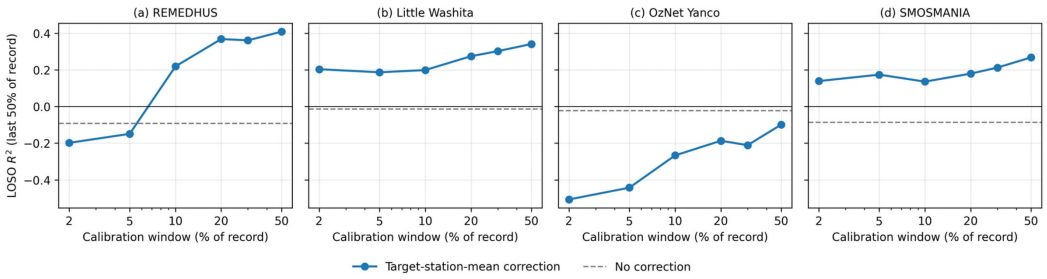



**Figure 10.** Sensitivity of target-station-mean correction to the calibration (warm-up) window length. LOSO $R^2$ (M3, canonical RF) of the corrected predictions (solid line) versus the uncorrected predictions (dashed line) for (**a**) REMEDHUS, (**b**) Little Washita, (**c**) OzNet Yanco, and (**d**) SMOSMANIA, as the warm-up window is increased from 2% to 50% of each held-out station's record. All windows are scored on the same last 50% of each station's record.

A further test asked whether the same station-mean information can be supplied as a model input instead of being used to shift the output after prediction (M4, Section 2.3; Table 10). Giving the held-out station the training-pool mean (M4-pool), which carries no information about that station, did not consistently remove the failure ($R^2$ = +0.046, −0.118, and +0.027 at REMEDHUS, Little Washita, and SMOSMANIA, below the climatology baseline at the latter two; +0.007 at OzNet). Giving it the mean of its earliest 10% of observations (M4-warm), the same information as the target-mean correction, raised $R^2$ to +0.390, +0.420, and +0.238 at REMEDHUS, Little Washita, and SMOSMANIA, higher than both the uncorrected values and the post hoc target-mean correction (+0.291, +0.332, +0.159) and above the climatology baseline, using a single model per network. At OzNet, M4-warm gave +0.031 (uncorrected: +0.041), i.e., no benefit, and even supplying the station's own full-record mean (not deployable, reported only as an upper bound) gave −0.006. The station-baseline shift is therefore learnable in feature space when information on the target station's mean is available, but not without it, and the output-shifting form of Section 3.7 is not the only remedy. These are single LOSO runs without confidence intervals, so differences of about 0.1 between M4-warm and the post hoc correction should be read as indicative only.

**Table 10.** Station-Mean Feature Model (M4) versus target-mean correction (Canonical RF, LOSO $R^2$, later 90% of each held-out station).

| Region | No Correction | Target-Mean Corr. | M4-Pool | M4-Warm | Climatology |
|---|---|---|---|---|---|
| REMEDHUS | −0.062 | +0.291 | +0.046 | +0.390 | +0.032 |
| Little Washita | −0.023 | +0.332 | −0.118 | +0.420 | +0.024 |
| OzNet Yanco | +0.041 | −0.132 | +0.007 | +0.031 | +0.096 |
| SMOSMANIA | −0.055 | +0.159 | +0.027 | +0.238 | +0.126 |

Note: M4-pool: held-out station receives the training-pool mean; M4-warm: held-out station receives the mean of its earliest 10% of observations.

Sections 3.1–3.7 assemble a multi-step diagnostic chain, in which each result narrows the set of candidate explanations and motivates the next test. Figure 11 summarizes this logic as a single roadmap, spanning from the initial accuracy question (Section 3.1) to the main conclusions (Section 5), before the Discussion interprets each step in detail. Because the target-mean correction (and M4-warm) requires in situ observations at the target station, it does not demonstrate generalization to completely unobserved locations.

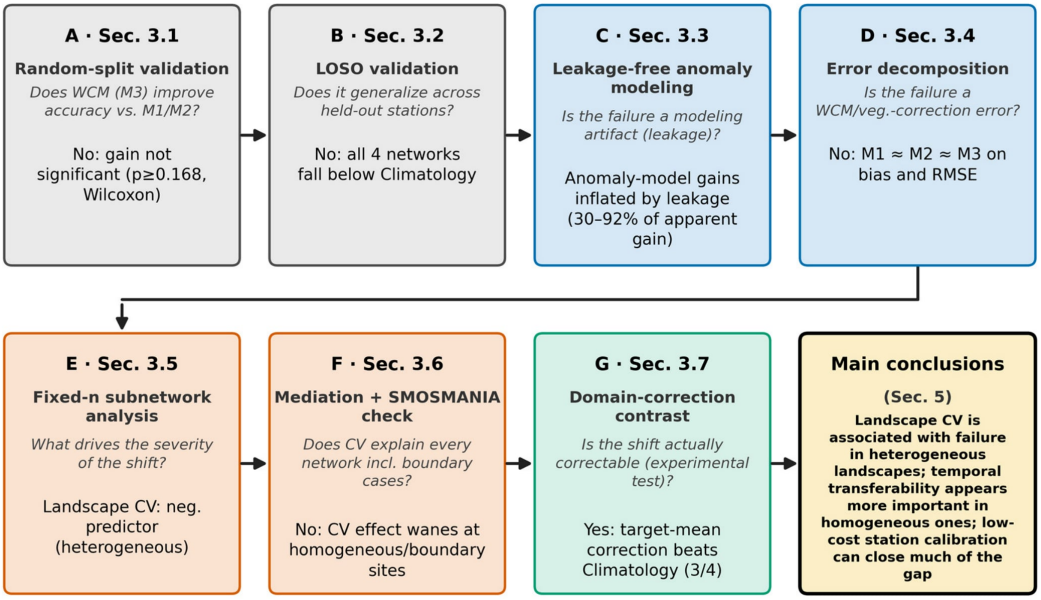



**Figure 11.** Diagnostic-protocol flowchart summarizing the multi-step logic of Sections 3.1–3.7. Top row (**A**–**D**): the initial accuracy/generalization question and two checks that examine whether the failure reflects a modeling artifact or a WCM-specific error. Bottom row (**E**–**G**): the mechanism (landscape heterogeneity), its boundary cases, and an experimental test of correctability, leading to the main conclusions stated in Section 5.

## 4. Discussion

### *4.1. Likely Cause of Cross-Site Generalization Failure: Distributional Shift as the Most Consistent Explanation*

The LOSO and error-decomposition results (Sections 3.2 and 3.4) indicate that the failure is more plausibly attributable to distributional shift than to the accuracy of vegetation correction (covariate shift [68]): the SAR–SM relationship differs systematically across stations, and the mapping function learned by RF on the training set cannot extrapolate to new stations, regardless of whether the WCM-decoupled feature is included. Section 3.7 provides supporting evidence for this diagnosis, including a test in which the same station-mean information is supplied as a model input (M4) rather than as a post hoc shift. These results show that distributional shift is sufficient to account for, and consistent with, the failure, but not that it is the only cause: point-to-pixel scale mismatch, sensor and label noise, and residual differences in preprocessing among networks may also contribute and cannot be separated with the present design (see Section 4.5 and the limitations).

This empirical chain admits a simple statistical explanation that does not require additional data to test. Consider a network of N stations with true long-term mean soil moisture $\mu_1, \ldots, \mu_N$ and between-station standard deviation $\sigma_\mu$, so that the landscape coefficient of variation (CV) is given by

$$\mathrm{CV} \approx \sigma_\mu / \mathrm{mean}(\mu) \quad (4)$$

A LOSO-trained RF has no station-identity feature and therefore implicitly calibrates its output to the training pool's aggregate baseline; when station k is held out, the expected magnitude of the resulting station-mean bias is $|\mu_k - \text{mean of the remaining } N-1 \text{ stations}|$, whose variance, for moderate-to-large N, is approximately

$$\mathrm{Var}(\mathrm{bias}_k) \approx \sigma_\mu^2 \cdot N/(N-1) \approx \sigma_\mu^2 \quad (5)$$

and therefore scales with CV. Under the standard bias–variance identity for out-of-sample $R^2$,

$$R^2 \approx 1 - (\text{Bias}^2 + \text{residual variance})/\text{Var}(y) \quad (6)$$

where y denotes the observed (in situ) soil moisture, consistent with SM_insitu used elsewhere in this paper; a larger expected station-mean bias mechanically depresses $R^2$, independently of how well the temporal SAR–SM dynamics are captured. This gives an analytic account of why LOSO $R^2$ declines with landscape CV (Section 3.5), why a substantial station-mean bias persists in the LOSO error decomposition regardless of the feature set (Section 3.4), and why a target-station-mean correction succeeds where feature-level or sample-level corrections do not (Section 3.7): it directly, and with low variance, estimates exactly the bias term this argument identifies as dominant, whereas the latter two corrections do not target this term at all. Supplying the same mean as a model input (M4, Section 3.7) likewise targets this term and likewise succeeds at the three heterogeneous networks. The argument is agnostic to whether WCM decoupling is used, consistent with the M1 ≈ M2 ≈ M3 equivalence noted in Section 3.4.

### *4.2. Station-Mean SM Heterogeneity (CV): An Effective Predictor, but with Boundaries*

As shown in Section 3.5, LOSO $R^2$ decreases monotonically with CV across the four networks, but the CV–$R^2$ relationship at the subnetwork level is robust only in the most heterogeneous network (REMEDHUS). Taken together, CV is an effective empirical predictor of cross-site generalization, but its scope has boundaries: the relationship is pronounced in highly heterogeneous regions and weakens or disappears in moderately heterogeneous or homogeneous regions, a boundary repeatedly observed across two independent, low-to-moderate-heterogeneity networks (Little Washita and SMOSMANIA). Because CV is computed from station-mean SM, it also reflects climatic, soil-texture, and hydrological differences among stations, so the association should be read as one with this composite heterogeneity rather than with landscape structure alone; the contrasting leakage-free behavior of SMOSMANIA and Little Washita, which have similar CV (Section 3.3), further shows that CV does not capture all relevant factors.

### *4.3. OzNet: Temporal Transferability Dominates in a Homogeneous Landscape*

In contrast to REMEDHUS and Little Washita, leakage-free anomaly modeling retains a substantial improvement at OzNet (Section 3.3). This suggests that, once station-baseline information is removed, OzNet's temporal dynamics (the covariation between SM and SAR/NDVI) can transfer from training stations to the test station. In a homogeneous landscape, inter-station baseline differences are small, distributional shift is no longer the primary obstacle, and temporal transferability becomes the key factor determining LOSO performance. This is consistent with OzNet's improvement under anomaly modeling being 'genuine,' whereas the improvements at REMEDHUS/Little Washita are almost entirely due to leakage. SMOSMANIA provides an important counterexample to this narrative (Section 3.3): its CV is nearly identical to that of Little Washita, yet its leakage-free improvement is close to that of OzNet. This suggests that CV alone does not determine the strength of the leakage-free transferable signal—factors such as inter-station climatic consistency, land-cover heterogeneity, or network density may likewise modulate the transferability of temporal signals, and CV alone is insufficient to fully characterize this. This complexity suggests that the CV–transferability relationship warrants further testing with additional networks rather than being reduced to a simple univariate framework; see Section 4.7 for further discussion.

### *4.4. Methodological Caution Regarding Anomaly Modeling*

By applying a leakage-free correction (Section 3.3), this study recovers the true magnitude of improvement, indicating that anomaly-modeling frameworks under LOSO evaluation should separate training-station information from test-station information, otherwise the reported gains may be substantially inflated. This type of performance inflation, arising from incomplete isolation of training/test information, is consistent with the machine-learning leakage patterns systematically cataloged in [41].

### *4.5. Precisely Locating the Contribution of WCM*

Within the RF-LOSO framework (Section 3.4), VV_soil as a single additional feature cannot overcome the distributional-shift bottleneck. WCM's contribution may only become apparent within a feature-engineering framework (spatializing the B parameter) or a domain-adaptation framework [69]. At OzNet and SMOSMANIA, where B could not be calibrated to a clear optimum (Section 2.2.3), VV_soil is likewise not informative as an RF feature.

Comparability of WCM parameterization across networks. The WCM was not parameterized identically across the four networks: REMEDHUS used region-specific literature A parameters (A_VV = 0.0046, A_VH = 0.0017), whereas the other networks shared A_VV = 0.10 and A_VH = 0.05; B was calibrated by grid search at REMEDHUS and Little Washita, whereas at OzNet and SMOSMANIA, where no clear optimum was found, the literature values were used (Section 2.2). The decoupled feature VV_soil is therefore not constructed under strictly identical conditions across networks, and cross-network differences in WCM effectiveness should be interpreted with this in mind. Four results nevertheless indicate that these differences do not drive the main findings. First, the incremental LOSO effect of the WCM term (M3 minus M2) is negligible at every network (+0.008, +0.001, −0.001, and +0.002 at REMEDHUS, Little Washita, OzNet, and SMOSMANIA, respectively; Table 5), including at REMEDHUS with its region-specific A and at OzNet and SMOSMANIA with their uncalibrated B. Second, station-mean bias and temporal-dynamic RMSE are virtually identical among M1, M2, and M3 at all networks (Table 7), so the diagnosis of distributional shift does not depend on the WCM term. Third, re-calibrating B within each LOSO fold from the training stations only changed the M3 $R^2$ by at most 0.0017 (Section 2.2.3). Fourth, omitting the vegetation-scattering term altogether ($\sigma$_soil = $\sigma$_total/$T^2$) or replacing it with a simplified form ($\sigma$_veg = A·$NDVI^2$) changed the M3 LOSO $R^2$ by at most 0.007 at any network (Supplementary Table S3). Differences in parameterization are therefore unlikely to account for the cross-network differences reported here, although a fully harmonized parameterization (for example, shared A parameters and spatially varying B) would be needed to exclude residual effects. Validation protocols likewise differed in scale among the networks (station counts of 12–21 and different numbers of bootstrap resamples), so cross-network comparisons of absolute $R^2$ should be read as comparisons of qualitative patterns, not as strictly controlled contrasts.

### *4.6. Comparison with Existing Literature*

Existing WCM-ML literature reports accuracies such as $R^2$ = 0.49–0.88 [19] and RMSE = 4.0–6.1%vol [18] under random or field-level splitting, which are not directly comparable to, and generally better than, this study's random-validation results (Table 4: $R^2$ = 0.03–0.28); a recent global-scale deep-learning data-fusion model reported an average per-sensor correlation of about 0.73 on held-out sensors and about 0.71 when the test sensors were required to lie more than 25 km from any training sensor, across roughly 1000 stations [21], i.e., only a small degradation with spatial separation. A more recent line of work instead couples WCM and deep networks end-to-end in a differentiable frame-

work [70], illustrating a research direction complementary to the diagnostic, validation-protocol focus of the present study: tighter physical–ML integration; whether such tighter coupling is itself subject to the same LOSO failure mode diagnosed here is an open question for future cross-study comparison. However, LOSO validation reveals that these results largely reflect 'within-station memory': RF falls below the monthly mean baseline under LOSO, and the leakage-correction experiments show that most of the apparent improvement from anomaly modeling stems from a methodological flaw. The extreme divergence between random and LOSO performance (e.g., REMEDHUS: 0.104 vs. −0.064) highlights the importance of LOSO in assessing real-world deployment value, consistent with the argument that spatial prediction models must have their area of applicability evaluated separately, outside the training domain [34,35,71]. More broadly, soil moisture is not only a direct target of remote sensing retrieval but also a core state variable in Earth-system-science questions such as land–atmosphere coupling [72] and global drought monitoring [73–75]; the integration of deep learning with process understanding is emerging as a new paradigm for data-driven geoscience research [76,77], which is also the broader context motivating this study's choice to incorporate a machine-learning component atop the physical WCM framework.

Table 11 situates this study's random-split and LOSO results against six representative WCM- or ML-based soil-moisture retrieval studies. Most works reporting high accuracy under random or field-level splits ([18–20]) do not test a genuine spatial holdout. Among the studies that do test spatial/cross-site generalization ([21,70,78]), the reported degradation from within-sample to out-of-sample performance is comparatively modest (e.g., $\Delta r \approx 0.02$ in [21]; $\Delta R \approx 0.06$ in [70]), in contrast to the collapse to below-Climatology performance observed under LOSO in the present study. This contrast plausibly reflects a difference in network scale and diversity: the large, geographically dispersed networks used in [78] and [21] (461–1000 stations spanning many climates) may average out station-level idiosyncrasies, whereas this study's four dense, internally heterogeneous regional networks (Table 1) more directly probe within-network distributional shift—precisely the regime in which the applicability boundary of WCM-ML is diagnosed here.

**Table 11.** Comparison with representative WCM- and ML-based soil moisture retrieval studies.

| Ref. | SAR/Data & Retrieval Method | Networks/Stations | Validation Scheme | Reported Accuracy | Explicit Spatial/Cross-Site Test? |
|---|---|---|---|---|---|
| [18] | S1 + S2; neural-network inversion of WCM + IEM | 23 field plots (France) | NN trained on synthetic data; tested on real in situ obs. | RMSE = 4.0–6.1%vol ($R^2$ n.r.) | No |
| [19] | ENVISAT/RADARSAT-2; ANN + WCM/NDVI | 6 sites (Italy/Australia/Spain) | ANN trained on simulated + experimental data; tested on separate real dataset | $R^2$ = 0.49–0.88 (config-dependent) | No |
| [20] | S1; WCM (vegetated) + linear regression (bare soil) | 7 stations + field campaign | Field-level split (45 calibration/45 validation fields) | Backscatter RMSE = 1.2–1.3 dB (SM% n.r.) | Partial (field split only) |
| [78] | S1 + Landsat-8 + GLDAS; Gradient-Boosted Regression Trees | 461 ISMN stations, 13 networks | Leave-One-Group-Out CV + independent 20% holdout | $R^2$ = 0.81 (holdout); median per-station temporal $R^2$ = 0.48 | Yes (LOGO-CV) |
| [21] | S1 + S2 + SoilGrids + SMAP + GLDAS; deep-learning fusion (CNN) | ~1000 global stations | IID sensor-level split vs. "hard" split (>25 km from the nearest training sensor) | r = 0.727 (IID) vs. 0.707 (spatial) | Yes (explicit spatial-holdout comparison) |

**Table 11.** *Cont*.

| Ref. | SAR/Data & Retrieval Method | Networks/Stations | Validation Scheme | Reported Accuracy | Explicit Spatial/Cross-Site Test? |
|---|---|---|---|---|---|
| [70] | S1; differentiable DNN–WCM hybrid | 4 regions (station count n.r.) | Within-site testing vs. "extrapolation" test | R = 0.75–0.85 (within-site) vs. 0.79 (extrapolation) | Yes (explicit extrapolation test) |
| This study | S1 + S2; WCM decoupling (M1/M2/M3) + Random Forest | 4 dense regional networks (Table 1) | Random-split Bootstrap vs. LOSO | Random $R^2$ = 0.03–0.28 (M1–M3); LOSO $R^2$ = −0.159 to +0.037 (M1–M3; below Climatology) | Yes (LOSO, with explicit random-vs-spatial contrast) |

Note: S1 = Sentinel-1; S2 = Sentinel-2; NN/ANN = (artificial) neural network; CNN = convolutional neural network; DNN = deep neural network; n.r. = not reported in the original publication; IEM = integral equation model; IID = independent and identically distributed. Values for [19] are reported per sensor/polarization configuration in the original study; the range shown spans its lowest- to highest-accuracy configuration. "Explicit spatial/cross-site test" indicates whether the cited study evaluated performance on stations or locations withheld from training (leave-one-station/group-out, or a train–test split enforced by a minimum spatial separation), as opposed to a random split of pooled, mixed-station data.

### *4.7. Limitations and Future Work*

Limitations: (1) a network of four sites is still insufficient to establish a quantitative law; the similar CV of SMOSMANIA and Little Washita provides cross-climate replication for one CV level but does not yet span the full heterogeneity spectrum; in addition, CV is derived from station-mean SM and therefore also reflects climatic, soil, and hydrological differences, and no independent landscape metric (e.g., land-cover diversity or soil-texture variability) was tested, so the heterogeneity effect cannot be attributed to landscape structure alone; (2) the fixed-n conclusion at Little Washita, like SMOSMANIA and OzNet, shows a significant correlation under lightweight RF that does not replicate under canonical RF, and N = 40 has limited statistical power to fully resolve this; (3) regional-mean substitution in leakage-free modeling introduces some error; (4) there is incomplete comparability between the large-scale, lightweight-RF fixed-n results and the small-sample canonical-RF results, the latter having limited statistical power at N = 40; (5) within the SMOSMANIA network, the Pézenas and Prades-le-Lez stations are only about 270 m apart (both are officially recognized independent stations), so LOSO leave-one-out may carry a slight risk of station non-independence between these two folds, which was not specifically addressed; (6) the target-mean correction in Section 3.7 relies on a brief in situ calibration period at each newly deployed station; although window lengths of 2–50% were tested (Figure 10), its sensitivity to noisier warm-up data and its applicability to fully ungauged sites remain to be tested; (7) in situ probes measure soil moisture at point scale, whereas the Sentinel-1 value represents a pixel or a small buffer, so part of the station-mean bias may reflect point-to-pixel mismatch; repeating the main analyses with 10 m, 30 m, and 100 m extraction windows (Supplementary Section S4) did not change the below-climatology LOSO performance, the station-mean bias, or the direction of the target-mean correction effect, but this mismatch cannot be removed with the present data. (8) The subnetworks in the fixed-n analysis overlap and are not independent, so the associated *p*-values are descriptive, and the preprocessing and validation procedures were not identical across networks. Future work: validate the CV–LOSO relationship using additional networks such as the Soil Climate Analysis Network (SCAN) and COSMOS-Europe [79] integrated via ISMN [10]; extend the target-mean correction validated in Section 3.7, which succeeded where CORAL-style feature standardization and few-shot fine-tuning did not, to more principled domain-adaptation methods [69,80,81]; investigate predicting station climatological baselines from SAR covariates to improve leakage-free modeling; and further test

network characteristics beyond CV (such as climatic consistency and network density) that may modulate the strength of the temporally transferable signal. Future work could also draw on transfer-learning theory [82,83] and deep domain-adaptation methods [80,84] to systematically model cross-site distributional shift; satellite soil moisture downscaling techniques [85,86] and deep-learning-based temporal modeling methods [87,88], together with the physical foundations of passive-microwave retrieval [89], all offer potential technical pathways for improving the cross-site generalization ability of WCM-ML.

### *4.8. Preliminary Recommendations for Deployment at New, Unseen Sites Based on the Networks Examined*

The diagnostic results in Sections 3 and 4.1, Sections 4.2–4.7 motivate a practical question that a diagnosis alone does not answer: given a newly instrumented or ungauged site, how should a practitioner decide whether a WCM-ML model's reported accuracy can be trusted, and what corrective action, if any, should be applied? Table 12 reorganizes this study's own results—the landscape-heterogeneity (CV) ordering, the uncorrected LOSO performance, and the target-station-mean-corrected performance already reported in Tables 1, 5 and 9—into a single reference point for reasoning about a new site of comparable heterogeneity; it introduces no new results.

**Table 12.** Empirical basis for the preliminary deployment recommendations: landscape heterogeneity, uncorrected LOSO performance, and target-mean-corrected performance.

| Network (Reference) | Landscape CV | LOSO $R^2$ (No Correction) | Vs. Climatology Baseline | LOSO $R^2$ (Target-Mean Correction, 10% Warm-Up) |
|---|---|---|---|---|
| REMEDHUS | 0.673 (high) | −0.064 | Below (Clim = +0.032) | +0.291 |
| SMOSMANIA | 0.380 (moderate) | −0.044 | Below Climatology | +0.159 |
| Little Washita | 0.364 (moderate) | −0.024 | Below Climatology | +0.332 |
| OzNet Yanco | 0.184 (low) | +0.036 | Below (Clim = +0.096) | −0.132 (harmful) |

Note: CV = coefficient of variation in multi-year mean soil moisture across stations (Table 1), used here as a proxy for landscape/network heterogeneity. Target-mean correction shifts each held-out station's predictions using only the mean of that station's earliest 10% of observations, simulating a brief calibration period at deployment (Section 3.7). Feature standardization (CORAL-style) and few-shot fine-tuning, tested under the same protocol, provided little or no improvement at any of the four networks and are omitted here for brevity; see Table 9 for the full five-setting comparison. Uncorrected values are from Table 5 (full record) and corrected values are from Table 9 (later 90% of each record), so the two columns are not scored on identical samples; like-for-like uncorrected values are given in Table 9.

Based on this evidence, three deployment scenarios can be distinguished. (a) If the new site's landscape CV can be estimated (e.g., from a short reconnaissance period or nearby reference stations) and is moderate-to-high (comparable to SMOSMANIA, Little Washita, or REMEDHUS, CV $\gtrsim 0.36$), and an in situ warm-up period covering at least about 10% of the station's record is available at deployment (Figure 10), target-station-mean correction is a reasonable option to consider: within this study it recovered positive, above-Climatology LOSO performance at all three such networks (+0.16 to +0.33), whereas feature standardization and few-shot fine-tuning did not. (b) If the new site's landscape is comparatively homogeneous (CV $\lesssim 0.2$, comparable to OzNet), target-mean correction may not be worthwhile by default; it was harmful at OzNet in this study ($R^2 = -0.132$ versus +0.041 without correction under the same scoring window, $\Delta R^2 = -0.173$), and effort is better directed at verifying the temporal-dynamic signal quality, since Section 4.3 indicates that temporal transferability, rather than correction, is the dominant lever in homogeneous landscapes. (c) If no in situ warm-up data are available at all (a fully ungauged site), none

of the corrections tested here apply, and an uncorrected WCM-ML model is unlikely to be reliable on the strength of its random-split accuracy: every network in this study saw its uncorrected LOSO $R^2$ fall below the simple Climatology/persistence baseline (Table 5), so the baseline itself may be the more defensible choice until in situ calibration data become available. Conversely, where a station's own antecedent in situ record is available, supplying lagged soil moisture as a predictor raised the mean held-out-station $R^2$ from −0.60 to 0.51 at Little Washita in our earlier work, although the zero-shot transfer to REMEDHUS was attributed mainly to soil-moisture persistence rather than to transferred SAR–optical relationships [39].

These preliminary recommendations are grounded in four networks and should be treated as a working hypothesis rather than a validated operational rule; as noted in Section 4.7, extending it to a broader range of networks (e.g., via ISMN [10]) remains an important direction for future work.

## 5. Conclusions

Through multidimensional experiments on four validation networks, this study diagnoses the failure of WCM-ML under cross-site (LOSO) generalization and draws the following conclusions:

(1) The WCM-decoupled feature yields no significant gain under random validation ($p \geq 0.168$; $|\Delta R^2| \leq 0.003$), nor in LOSO $R^2$, station-mean bias, or temporal-dynamic error (M1 ≈ M2 ≈ M3). Random-split validation therefore overestimates the practical value of WCM, and under LOSO all models fall below the monthly climatology baseline, a failure most consistent with a shift in each station's absolute SM baseline rather than with vegetation-correction error.

(2) LOSO $R^2$ decreases monotonically with landscape heterogeneity (CV) across the four networks, but at the subnetwork level CV is a robust negative predictor only in the most heterogeneous network (REMEDHUS); the relationship weakens or disappears in moderately heterogeneous and homogeneous networks, and the positive correlation at OzNet under lightweight RF does not replicate under canonical RF. SMOSMANIA and Little Washita, with similar CV, provide cross-climate replication of this boundary.

(3) The large apparent gains of anomaly modeling are largely artifacts of station-mean leakage at REMEDHUS and Little Washita (86% and 92% of the apparent gain); only OzNet and SMOSMANIA retain a meaningful leakage-free improvement. That SMOSMANIA differs from Little Washita despite similar CV indicates that CV alone does not determine the transferable temporal signal.

(4) A target-station-mean correction calibrated from only the earliest 10% of a new station's observations lifts LOSO $R^2$ above the climatology baseline at REMEDHUS, Little Washita, and SMOSMANIA, whereas feature standardization and few-shot fine-tuning provide little benefit, supporting the interpretation that the failure is a correctable station-baseline shift. The same station-mean information supplied as a model input (M4) gives comparable or larger gains with a single model ($R^2$ = +0.39, +0.42, and +0.24 at REMEDHUS, Little Washita, and SMOSMANIA), but largely only when the target station's own mean is available.

(5) At OzNet and SMOSMANIA, the WCM B parameter could not be calibrated to a clear interior optimum; the main conclusions were insensitive to this and to the treatment of the vegetation-scattering term.

Taken together, the evidence from the four networks examined suggests that landscape heterogeneity (CV) is associated with lower WCM-ML cross-site generalization across the four networks, although the subnetwork-level relationship is robust only in the most heterogeneous landscape and wanes as the landscape becomes more homogeneous; in

homogeneous landscapes, temporal transferability, rather than the level of heterogeneity, appears to be the key factor. Improving cross-site generalization in heterogeneous landscapes may require domain adaptation, spatially varying B, or better observation-network design, although a low-cost deployment-time station calibration (Section 3.7) may already close much of the gap in the networks tested.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs18213842/s1, Section S1: Data processing details; Section S2: History of the B-parameter grid search; Section S3: Sensitivity of the results to the vegetation-scattering term; Section S4: Sensitivity to the extraction window.

**Author Contributions:** Conceptualization, P.W. and Q.J.; methodology, P.W.; software, P.W.; validation, P.W.; formal analysis, P.W.; investigation, P.W.; resources, Q.J.; data curation, P.W.; writing—original draft preparation, P.W.; writing—review and editing, C.Y. and Q.J.; visualization, P.W.; supervision, Q.J.; project administration, Q.J.; funding acquisition, Q.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the China Geological Survey Project, grant number DD20191011.

**Data Availability Statement:** The in situ soil moisture data used in this study were derived from the following resources available in the public domain: the International Soil Moisture Network (https://ismn.earth), and the REMEDHUS, SMOSMANIA, Little Washita, and OzNet monitoring networks. Sentinel-1/Sentinel-2 SAR and optical data were derived from the Copernicus Open Access Hub, accessed via Google Earth Engine (https://earthengine.google.com). No new data repository was created for this study; the processing workflow required to reproduce the derived datasets is described in Section 2. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Petropoulos, G.P.; Ireland, G.; Barrett, B. Surface soil moisture retrievals from remote sensing: Current status, products & future trends. *Phys. Chem. Earth Parts A/B/C* **2015**, *83–84*, 36–56. [CrossRef]
2. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.-P.; Delwart, S.; Cabot, F.; Boutin, J.; Escorihuela, M.-J.; Font, J.; Reul, N.; Gruhier, C.; et al. The SMOS mission: New tool for monitoring key elements of the global water cycle. *Proc. IEEE* **2010**, *98*, 666–687. [CrossRef]
3. Entekhabi, D.; Njoku, E.G.; O'Neill, P.E.; Kellogg, K.H.; Crow, W.T.; Edelstein, W.N.; Entin, J.K.; Goodman, S.D.; Jackson, T.J.; Johnson, J.; et al. The Soil Moisture Active Passive (SMAP) mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
4. Li, X.; Wigneron, J.-P.; Frappart, F.; De Lannoy, G.; Fan, L.; Zhao, T.; Gao, L.; Tao, S.; Ma, H.; Peng, Z.; et al. The first global soil moisture and vegetation optical depth product retrieved from fused SMOS and SMAP L-band observations. *Remote Sens. Environ.* **2022**, *282*, 113272. [CrossRef]
5. Bauer-Marschallinger, B.; Freeman, V.; Cao, S.; Paulik, C.; Schaufler, S.; Stachl, T.; Modanesi, S.; Massari, C.; Ciabatta, L.; Brocca, L.; et al. Toward global soil moisture monitoring with Sentinel-1: Harnessing assets and overcoming obstacles. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 520–539. [CrossRef]
6. Sánchez, N.; Martínez-Fernández, J.; Scaini, A.; Pérez-Gutiérrez, C. Validation of the SMOS L2 soil moisture data in the REMEDHUS network (Spain). *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1602–1611. [CrossRef]
7. Cosh, M.H.; Jackson, T.J.; Starks, P.; Heathman, G. Temporal stability of surface soil moisture in the Little Washita River watershed and its applications in satellite soil moisture product validation. *J. Hydrol.* **2006**, *323*, 168–177. [CrossRef]
8. Smith, A.B.; Walker, J.P.; Western, A.W.; Young, R.I.; Ellett, K.M.; Pipunic, R.C.; Grayson, R.B.; Siriwardena, L.; Chiew, F.H.S.; Richter, H. The Murrumbidgee soil moisture monitoring network data set. *Water Resour. Res.* **2012**, *48*, W07701. [CrossRef]
9. Calvet, J.-C.; Fritz, N.; Froissard, F.; Suquia, D.; Petitpa, A.; Piguet, B. In situ soil moisture observations for the CAL/VAL of SMOS: The SMOSMANIA network. In Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–28 July 2007; pp. 1196–1199. [CrossRef]
10. Dorigo, W.; Himmelbauer, I.; Aberer, D.; Schremmer, L.; Petrakovic, I.; Zappa, L.; Preimesberger, W.; Xaver, A.; Annor, F.; Ardö, J.; et al. The International Soil Moisture Network: Serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 5749–5804. [CrossRef]
11. Attema, E.P.W.; Ulaby, F.T. Vegetation modeled as a water cloud. *Radio Sci.* **1978**, *13*, 357–364. [CrossRef]

12. Baghdadi, N.; El Hajj, M.; Zribi, M.; Bousbih, S. Calibration of the water cloud model at C-band for winter crop fields and grasslands. *Remote Sens.* **2017**, *9*, 969. [CrossRef]
13. Das, D.P.; Pandey, A. Soil moisture retrieval from dual-polarized Sentinel-1 SAR data over agricultural regions using a water cloud model. *Environ. Monit. Assess.* **2025**, *197*, 52. [CrossRef] [PubMed]
14. Ma, C.; Wang, S.; Zhao, Z.; Ma, H. Global sensitivity analysis of a water cloud model toward soil moisture retrieval over vegetated agricultural fields. *Remote Sens.* **2021**, *13*, 3889. [CrossRef]
15. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
16. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
17. Hou, C.; Tan, M.L.; Yaseen, Z.M.; Zhang, F. Advancing soil moisture retrieval from SAR data using meta-heuristic optimization XGBoost and SHAP analysis. *Int. J. Remote Sens.* **2025**, *46*, 5354–5383. [CrossRef]
18. El Hajj, M.; Baghdadi, N.; Zribi, M.; Bazzi, H. Synergic use of Sentinel-1 and Sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas. *Remote Sens.* **2017**, *9*, 1292. [CrossRef]
19. Paloscia, S.; Pettinato, S.; Santi, E.; Notarnicola, C.; Pasolli, L.; Reppucci, A. Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation. *Remote Sens. Environ.* **2013**, *134*, 234–248. [CrossRef]
20. Bousbih, S.; Zribi, M.; Lili-Chabaane, Z.; Baghdadi, N.; El Hajj, M.; Gao, Q.; Mougenot, B. Potential of Sentinel-1 radar data for the assessment of soil and cereal cover parameters. *Sensors* **2017**, *17*, 2617. [CrossRef] [PubMed]
21. Batchu, V.; Nearing, G.; Gulshan, V. A deep learning data fusion model using Sentinel-1/2, SoilGrids, SMAP, and GLDAS for soil moisture retrieval. *J. Hydrometeorol.* **2023**, *24*, 1789–1823. [CrossRef]
22. Zhang, R.; Song, C.; Zhang, Z.; Xie, J.; Chen, T.; Xu, T. Machine learning for soil moisture analysis: A survey and emerging perspectives. *Int. J. Data Sci. Anal.* **2025**, *21*, 66. [CrossRef]
23. Fung, A.K.; Li, Z.; Chen, K.S. Backscattering from a randomly rough dielectric surface. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 356–369. [CrossRef]
24. Baghdadi, N.; King, C.; Chanzy, A.; Wigneron, J.-P. An empirical calibration of the integral equation model based on SAR data and soil moisture and roughness field measurements. *Int. J. Remote Sens.* **2002**, *23*, 4325–4340. [CrossRef]
25. Oh, Y.; Sarabandi, K.; Ulaby, F.T. Semi-empirical model of the ensemble-averaged differential Mueller matrix for microwave backscattering from bare soil surfaces. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1348–1355. [CrossRef]
26. Dubois, P.C.; van Zyl, J.; Engman, T. Measuring soil moisture with imaging radars. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 915–926. [CrossRef]
27. Zribi, M.; Dechambre, M. A new empirical model to retrieve soil moisture and roughness from C-band radar data. *Remote Sens. Environ.* **2003**, *84*, 42–52. [CrossRef]
28. Wagner, W.; Lemoine, G.; Rott, H. A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sens. Environ.* **1999**, *70*, 191–207. [CrossRef]
29. Jackson, T.J.; Schmugge, T.J. Vegetation effects on the microwave emission of soils. *Remote Sens. Environ.* **1991**, *36*, 203–212. [CrossRef]
30. Ulaby, F.T.; Moore, R.K.; Fung, A.K. *Microwave Remote Sensing: Active and Passive, Vol. III: From Theory to Applications*; Artech House: Norwood, MA, USA, 1986.
31. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [CrossRef]
32. Park, J.; Bindlish, R.; Bringer, A.; Horton, D.; Johnson, J.T. NISAR time-series ratio algorithm for soil moisture retrieval: Prelaunch evaluation with SMAPVEX12 field campaign data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12959–12968. [CrossRef]
33. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
34. Meyer, H.; Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [CrossRef]
35. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **2020**, *11*, 4540. [CrossRef] [PubMed]
36. Wang, Y.; Khodadadzadeh, M.; Zurita-Milla, R. Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *121*, 103364. [CrossRef]
37. Milà, C.; Ludwig, M.; Pebesma, E.; Tonne, C.; Meyer, H. Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls. *Geosci. Model Dev.* **2024**, *17*, 6007–6033. [CrossRef]

38. Hoppe, H.; Dietrich, P.; Marzahn, P.; Weiß, T.; Nitzsche, C.; von Lukas, U.F.; Wengerek, T.; Borg, E. Transferability of machine learning models for crop classification in remote sensing imagery using a new test methodology: A study on phenological, temporal, and spatial influences. *Remote Sens.* **2024**, *16*, 1493. [CrossRef]
39. Wang, P.; Jiang, Q. Sentinel-1 SAR and Temporal Lag Soil Moisture Estimation at Instrumented Field Sites: A Stacked Ensemble Approach. *Remote Sens.* **2026**, *18*, 2483. [CrossRef]
40. Tobler, W.R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef] [PubMed]
41. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
42. Vachaud, G.; Passerat De Silans, A.; Balabanis, P.; Vauclin, M. Temporal stability of spatially measured soil water probability density function. *Soil Sci. Soc. Am. J.* **1985**, *49*, 822–828. [CrossRef]
43. Western, A.W.; Grayson, R.B.; Blöschl, G. Scaling of soil moisture: A hydrologic perspective. *Annu. Rev. Earth Planet. Sci.* **2002**, *30*, 149–180. [CrossRef]
44. Famiglietti, J.S.; Ryu, D.; Berg, A.A.; Rodell, M.; Jackson, T.J. Field observations of soil moisture variability across scales. *Water Resour. Res.* **2008**, *44*, W01423. [CrossRef]
45. Ochsner, T.E.; Cosh, M.H.; Cuenca, R.H.; Hagimoto, Y.; Kerr, Y.H.; Njoku, E.G.; Zreda, M. State of the art in large-scale soil moisture monitoring. *Soil Sci. Soc. Am. J.* **2013**, *77*, 1888–1919. [CrossRef]
46. Mohanty, B.P.; Cosh, M.; Lakshmi, V.; Montzka, C. Remote sensing for vadose zone hydrology—A synthesis from the vantage point. *Vadose Zone J.* **2013**, *12*, 1–6. [CrossRef]
47. Topp, G.C.; Davis, J.L.; Annan, A.P. Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resour. Res.* **1980**, *16*, 574–582. [CrossRef]
48. Rodell, M.; Houser, P.R.; Jambor, U.; Gottschalck, J.; Mitchell, K.; Meng, C.-J.; Arsenault, K.; Cosgrove, B.; Radakovich, J.; Bosilovich, M.; et al. The Global Land Data Assimilation System. *Bull. Am. Meteorol. Soc.* **2004**, *85*, 381–394. [CrossRef]
49. Gruber, A.; Scanlon, T.; van der Schalie, R.; Wagner, W.; Dorigo, W. Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology. *Earth Syst. Sci. Data* **2019**, *11*, 717–739. [CrossRef]
50. Preimesberger, W.; Stradiotti, P.; Dorigo, W. ESA CCI Soil Moisture GAPFILLED: An independent global gap-free satellite climate data record with uncertainty estimates. *Earth Syst. Sci. Data* **2025**, *17*, 4305–4329. [CrossRef]
51. Zreda, M.; Shuttleworth, W.J.; Zeng, X.; Zweck, C.; Desilets, D.; Franz, T.; Rosolem, R. COSMOS: The cosmic-ray soil moisture observing system. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 4079–4099. [CrossRef]
52. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In Proceedings of the 3rd Earth Resources Technology Satellite-1 Symposium, NASA SP-351, Washington, DC, USA, 10–14 December 1973; pp. 309–317.
53. Shi, J.; Yang, H.; Hou, X.; Zhang, H.; Tang, G.; Zhao, H.; Wang, F. Coupling SAR and optical remote sensing data for soil moisture retrieval over dense vegetation covered areas. *PLoS ONE* **2025**, *20*, e0315971. [CrossRef] [PubMed]
54. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
55. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
56. Sun, J.; Ongsomwang, S. Optimal parameters of random forest for land cover classification with suitable data type and dataset on Google Earth Engine. *Front. Earth Sci.* **2023**, *11*, 1188093. [CrossRef]
57. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
58. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
59. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing System (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
60. Huang, F.; Zhang, Y.; Zhang, Y.; Nourani, V.; Li, Q.; Li, L.; Shangguan, W. Towards interpreting machine learning models for predicting soil moisture droughts. *Environ. Res. Lett.* **2023**, *18*, 074002. [CrossRef]
61. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331. [CrossRef]
62. McColl, R.H.; Vogelzang, J.; Konings, A.G.; Entekhabi, D.; Piles, M.; Stoffelen, A. Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophys. Res. Lett.* **2014**, *41*, 6229–6236. [CrossRef]
63. Crow, W.T.; Berg, A.A.; Cosh, M.H.; Loew, A.; Mohanty, B.P.; Panciera, R.; de Rosnay, P.; Ryu, D.; Walker, J.P. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. *Rev. Geophys.* **2012**, *50*, RG2002. [CrossRef]

64. Xie, Q.; Jia, L.; Menenti, M.; Hu, G. Global soil moisture data fusion by triple collocation analysis from 2011 to 2018. *Sci. Data* **2022**, *9*, 687. [CrossRef] [PubMed]
65. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
66. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
67. Rahmati, M.; Amelung, W.; Brogi, C.; Dari, J.; Flammini, A.; Bogena, H.; Brocca, L.; Chen, H.; Groh, J.; Koster, R.D.; et al. Soil moisture memory: State-of-the-art and the way forward. *Rev. Geophys.* **2024**, *62*, e2023RG000828. [CrossRef]
68. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **2000**, *90*, 227–244. [CrossRef]
69. Tuia, D.; Persello, C.; Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
70. Li, Z.; Yuan, Q.; Yang, Q.; Li, J.; Zhao, T. Differentiable modeling for soil moisture retrieval by unifying deep neural networks and water cloud model. *Remote Sens. Environ.* **2024**, *311*, 114281. [CrossRef]
71. Zhu, L.; Dai, J.; Jin, J.; Yuan, S.; Xiong, Z.; Walker, J.P. Are the current expectations for SAR remote sensing of soil moisture using machine learning overoptimistic? *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4501815. [CrossRef]
72. Koster, R.D.; Dirmeyer, P.A.; Guo, Z.; Bonan, G.; Chan, E.; Cox, P.; Gordon, C.T.; Kanae, S.; Kowalczyk, E.; Lawrence, D.; et al. Regions of strong coupling between soil moisture and precipitation. *Science* **2004**, *305*, 1138–1140. [CrossRef] [PubMed]
73. Sheffield, J.; Wood, E.F. Global trends and variability in soil moisture and drought characteristics, 1950–2000, from observation-driven simulations of the terrestrial hydrologic cycle. *J. Clim.* **2008**, *21*, 432–458. [CrossRef]
74. Qing, Y.; Wang, S.; Yang, Z.-L.; Gentine, P. Soil moisture–atmosphere feedbacks have triggered the shifts from drought to pluvial conditions since 1980. *Commun. Earth Environ.* **2023**, *4*, 254. [CrossRef]
75. Hirschi, M.; Stradiotti, P.; Crezee, B.; Dorigo, W.; Seneviratne, S.I. Potential of long-term satellite observations and reanalysis products for characterising soil drying: Trends and drought events. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 397–425. [CrossRef]
76. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
77. Zhu, X.X.; Xiong, Z.; Wang, Y.; Stewart, A.J.; Heidler, K.; Wang, Y.; Yuan, Z.; Dujardin, T.; Xu, Q.; Shi, Y. On the foundations of Earth foundation models. *Commun. Earth Environ.* **2026**, *7*, 103. [CrossRef]
78. Greifeneder, F.; Notarnicola, C.; Wagner, W. A machine learning-based approach for surface soil moisture estimations with Google Earth Engine. *Remote Sens.* **2021**, *13*, 2099. [CrossRef]
79. Bogena, H.R.; Schrön, M.; Jakobi, J.; Ney, P.; Zacharias, S.; Andreasen, M.; Baatz, R.; Boorman, D.; Duygu, M.B.; Eguibar-Galán, M.A.; et al. COSMOS-Europe: A European network of cosmic-ray neutron soil moisture sensors. *Earth Syst. Sci. Data* **2022**, *14*, 1125–1151. [CrossRef]
80. Xu, M.; Wu, M.; Chen, K.; Zhang, C.; Guo, J. The eyes of the gods: A survey of unsupervised domain adaptation methods based on remote sensing data. *Remote Sens.* **2022**, *14*, 4380. [CrossRef]
81. Zhu, L.; Cai, Q.; Jin, J.; Yuan, S.; Shen, X.; Walker, J.P. Multi-scale domain adaptation for high-resolution soil moisture retrieval from synthetic aperture radar in data-scarce regions. *J. Hydrol.* **2025**, *657*, 133073. [CrossRef]
82. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
83. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
84. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
85. Peng, J.; Loew, A.; Merlin, O.; Verhoest, N.E.C. A review of spatial downscaling of satellite remotely sensed soil moisture. *Rev. Geophys.* **2017**, *55*, 341–366. [CrossRef]
86. Senanayake, I.P.; Pathira Arachchilage, K.R.L.; Yeo, I.-Y.; Khaki, M.; Han, S.-C.; Dahlhaus, P.G. Spatial downscaling of satellite-based soil moisture products using machine learning techniques: A review. *Remote Sens.* **2024**, *16*, 2067. [CrossRef]
87. Fang, K.; Pan, M.; Shen, C. The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2221–2233. [CrossRef]
88. Singh, A.; Gaurav, K. Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images. *Sci. Rep.* **2023**, *13*, 2251. [CrossRef] [PubMed]
89. Njoku, E.G.; Entekhabi, D. Passive microwave remote sensing of soil moisture. *J. Hydrol.* **1996**, *184*, 101–129. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.