https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

A first 90 m resolution active layer moisture dataset across the Qinghai–Tibet Plateau permafrost region

Erji Du[1], Tonghua Wu[1], Liyun Dai[2], Youhua Ran[3], Yufang Min[3], Lin Zhao[4], Lingxiao Wang[4], Jimin Yao[1], Zhibin Li[4], Keting Feng[5], Yao Xiao[1], Guojie Hu[1], Defu Zou[1], Guangyue Liu[1], Zanpin Xing[6], Xiaofan Zhu[1], Ji Chen[7]

[1]Cryosphere Research Station on the Qinghai-Tibet Plateau, State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
[2]Heihe Remote Sensing Experimental Research Station, State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
[3]National Cryosphere Desert Data Center, State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 73000, China
[4]School of Geographical Sciences, Nanjing University of Information Science & Technology (NUIST), Nanjing 210044, China
[5]Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China
[6]Center for the Pan-Third Pole Environment, Lanzhou University, Lanzhou 730000, China
[7]Beiluhe Observation and Research Station of Frozen Soil Engineering and Environment, State Key Laboratory of Frozen Soil Engineering, Northwest Institute of Eco-Environmental and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

Correspondence to: Tonghua Wu (thuawu@lzb.ac.cn)

Abstract. Active layer moisture (ALM) plays a fundamental role in regulating the hydrothermal dynamics, freeze–thaw processes, carbon cycling, and ecosystem and load-bearing functions of permafrost. However, spatially continuous, high-resolution datasets that characterize soil moisture across the entire active layer remain largely unavailable for the Qinghai–Tibet Plateau (QTP), which hosts the largest low- to mid-latitude permafrost region globally. Here, we present the first 90 m resolution, spatially complete dataset of ALM—the depth-averaged volumetric water content of the entire active layer—for the entire permafrost region of the QTP. The dataset was produced by integrating 342 in situ samples collected during peak-thaw seasons from 2009 to 2024 with multi-source remote sensing environmental predictors, topographic factors, and soil physicochemical properties. Four ensemble machine learning models, include Random Forest, Extra Trees, XGBoost, and CatBoost, were trained, and a bias-aware multi-model fusion strategy was applied to generate the final ALM product. SHAP-based recursive feature elimination was used to identify the dominant environmental controls and optimize feature selection. Random five-fold cross-validation showed high apparent accuracy (R^2 =0.62–0.63, RMSE ≈ 0.08 $m^3 m^{-3}$), and a group-based spatial cross-validation provided a conservative transferability estimate (pooled R^2 = 0.30–0.38). Within the permafrost region, ALM ranges from 0.02 to 0.70 $m^3 m^{-3}$, with a mean of 0.21 $m^3 m^{-3}$ and a standard deviation of 0.09 $m^3 m^{-3}$, exhibiting a distinct southeast-to-northwest decreasing gradient. The dataset is freely available with DOI: https://doi.org/10.12072/ncdc.permafrost.db7312.2026. The dataset is accompanied by an area-of-applicability (AOA)

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

mask delineating the region—approximately 60% of the permafrost region—where the cross-validated performance can be expected to hold, together with per-pixel uncertainty estimates derived from the DI–RMSE relationship and the divergence among the four models.

1 Introduction

Permafrost regions represent a key component of the Earth system, exerting strong controls on land–atmosphere interactions, hydrological processes, and carbon cycling under a warming climate(IPCC, 2023; Jin et al., 2021; Schuur et al., 2015). The active layer, defined as the near-surface soil layer that undergoes seasonal freeze–thaw cycles, is the most dynamic interface mediating energy and mass exchange between the atmosphere and the underlying permafrost. The active layer thickness (ALT) has been recognized as an Essential Climate Variable (ECV) by the World Meteorological Organization, highlighting its importance in monitoring permafrost dynamics(GCOS, 2022). While ALT has been widely used as an indicator of permafrost change, increasing evidence suggests that active layer moisture (ALM) plays a more fundamental role in regulating hydrothermal processes(Hinkel and Nelson, 2003; Wlostowski et al., 2018). Throughout this paper, ALM specifically denotes the depth-averaged volumetric water content (VWC, $m^3\ m^{-3}$) of the entire active layer—from the ground surface down to the thaw front—rather than the moisture content at any individual depth; the two terms are therefore equivalent in this study, and ALM is used consistently as the single term throughout the manuscript. Because the full active-layer column can only be sampled once it has thawed to its annual maximum depth, ALM is observed at a well-defined point of the seasonal cycle, namely the peak-thaw season (September–October on the QTP), when the active layer is fully thawed and liquid water content reaches its annual maximum; the dataset presented here therefore characterises this peak-thaw seasonal state. Although soil moisture at individual depths varies considerably over the annual freeze–thaw cycle, the column-integrated nature of ALM renders its intra- and inter-annual variations relatively stable, so that peak-thaw-season surveys provide a representative characterisation of the annual ALM state. Specifically, ALM governs soil thermal properties, controls ALT and permafrost distribution, modulates water and energy fluxes(Du et al., 2023; Ji et al., 2022). It also directly influences vegetation growth and microbial activity, thereby linking physical processes with ecosystem and biogeochemical dynamics. Consequently, understanding the spatial distribution of ALM is essential for advancing our understanding of permafrost dynamics and their feedbacks to climate change.

The Qinghai–Tibet Plateau (QTP), often referred to as the “Asian Water Tower” (Yao et al., 2022), hosts the largest extent of permafrost in low- and mid-latitude regions and serves as the headwater area for many major Asian rivers. Permafrost degradation on the QTP—driven by ongoing atmospheric warming—has profound implications for regional hydrology, water security, and ecosystem stability(Cheng et al., 2019). Due to its complex topography and diverse climatic conditions, the spatial variability of active layer processes on the QTP is particularly pronounced. ALM plays a critical role in regulating active layer thickness and permafrost distribution, runoff generation, soil water storage, and surface energy balance across

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

the plateau. Therefore, characterizing its spatial distribution is not only crucial for understanding permafrost–hydrology interactions but also for assessing water resource dynamics and ecological responses in this climatically sensitive region.

Despite its importance, current knowledge of ALM remains limited by significant data and methodological constraints. First, in situ observations are sparse and unevenly distributed due to the harsh environmental conditions and logistical challenges of permafrost regions, resulting in insufficient spatial representativeness, particularly for subsurface measurements. Second, satellite-based soil moisture products—although providing broader spatial coverage—are generally limited to near-surface layers (< 5 cm), suffer from shallow sensing depth, coarse spatial resolution, and reduced accuracy under freeze–thaw conditions, leading to systematic biases in permafrost areas(Dorigo et al., 2017; Xing et al., 2025). While recent studies have produced high-resolution (100 m) surface soil moisture products for the QTP during the thawing season using Sentinel-1 SAR data (Li et al., 2025), the spatial distribution of soil moisture throughout the full depth of the active layer—which is critical for characterizing subsurface hydrothermal processes—remains largely unexplored. Consequently, there is still a critical lack of spatially continuous, high-resolution, and reliable datasets characterizing soil moisture throughout the entire active layer at regional scales. This data gap significantly hinders the accurate representation of hydrothermal processes and limits the predictive capability of land surface and permafrost models.

Machine learning approaches have emerged as powerful tools for spatial environmental prediction by capturing nonlinear relationships between environmental variables and target properties. Ensemble tree-based methods such as Random Forest (Breiman, 2001), Extra Trees (Geurts et al., 2006), XGBoost (Chen and Guestrin, 2016), and CatBoost (Dorogush et al., 2018) have demonstrated strong performance in soil property mapping tasks, particularly when integrated with multi-source remote sensing and terrain covariates (Hengl et al., 2017; Poggio et al., 2021; Shi et al., 2025). Furthermore, the SHAP (SHapley Additive exPlanations) interpretable machine learning framework (Lundberg and Lee, 2017) provides a theoretically grounded, model-agnostic approach to quantifying the relative importance of environmental drivers, thereby enhancing the physical interpretability of data-driven models and supporting scientific inference. The demonstrated capacity of these methods in global and regional soil property mapping offers considerable promise for high-accuracy spatial modelling of ALM across the QTP.

This study develops a high-resolution, spatially complete dataset of ALM across the permafrost region of the QTP by fusing long-term in situ observations with multi-source remote sensing, topographic, and soil physicochemical variables. The above four advanced ensemble tree-based machine learning models are adopted to construct robust nonlinear relationships between ALM and its key controlling factors. A bias-aware multi-model fusion strategy is applied to alleviate systematic biases arising from the uneven distribution of training samples. The SHAP-RFE (SHAP-based Recursive Feature Elimination) framework is integrated to quantify the relative importance of environmental drivers, optimize feature combinations, and enhance the physical interpretability and generalization ability of the models.

The key contributions of this study are fourfold: (1) It produces the first 90 m resolution, spatially continuous ALM dataset covering the entire permafrost region of the QTP; (2) It improves modeling accuracy and reliability by combining multi-model ensemble prediction with a bias-aware fusion strategy to alleviate systematic low-value overestimation and high-value

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

underestimation; (3) It quantitatively identifies the dominant controls of remote sensing, soil, and topographic factors on ALM using the SHAP interpretable machine learning framework; (4) It provides a publicly available, well-documented ALM product that supports permafrost hydrothermal modeling, hydrological simulation, ecological assessment, and climate change impact research.

2 Data and Methods

2.1 Field-Surveyed ALM Data

The field-surveyed ALM data were collected from multiple field campaigns conducted during the peak-thaw season (September–October), when the active layer has thawed to its annual maximum depth, across the permafrost region of the QTP between 2009 and 2024. The integration of sampling data from different years spanning 2009–2024 was motivated by two considerations. First, pooling multi-year samples maximises the training dataset size and mitigates the limited spatial coverage and insufficient sample density inherent to any single-year survey. Second, long-term monitoring records at representative QTP sites demonstrate that, although interannual fluctuations in active layer moisture do exist, their magnitude is negligible relative to the spatial variability across the study area and will therefore not introduce substantial bias into the spatial modelling results (seeSupplementary Fig. S1for details).

A total of 342 valid sample points were obtained, covering diverse permafrost types and land surface conditions. Their spatial distribution is shown in Fig. 1. Among these, 123 sample points were measured from active layer test pit profiles using the cutting ring (core sampling) method, with subsequent laboratory gravimetric determination of soil moisture; the remaining 219 sample points were derived from Ground-Penetrating Radar (GPR) surveys. The GPR-based approach，which was elaborated in Du et al. (2020)， first retrieves the electromagnetic wave propagation velocity within the active layer from common-midpoint profiles, then derives the apparent dielectric permittivity via the physical velocity–permittivity relationship (Eq. 1), and subsequently estimates ALM using an empirical permittivity–VWC conversion (Eq. 2). Equation 2 is a site-specifically calibrated formula developed for the QTP active layer based on the Complex Refractive Index Model (Huisman et al., 2003), with a maximum retrieval error in ALM of less than 0.08 $m^3\ m^{-3}$.

$$v = c/\sqrt{\varepsilon} \ , \tag{1}$$

$$\theta = 0.458\,\varepsilon^{0.26} - 0.664 \ , \tag{2}$$

where v is the electromagnetic wave propagation velocity (m ns^{-1}), c is the speed of light in vacuum (approximately 0.3 m ns^{-1}), ε is the relative dielectric permittivity of the active layer, and θ is the volumetric water content ($m^3\ m^{-3}$).

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

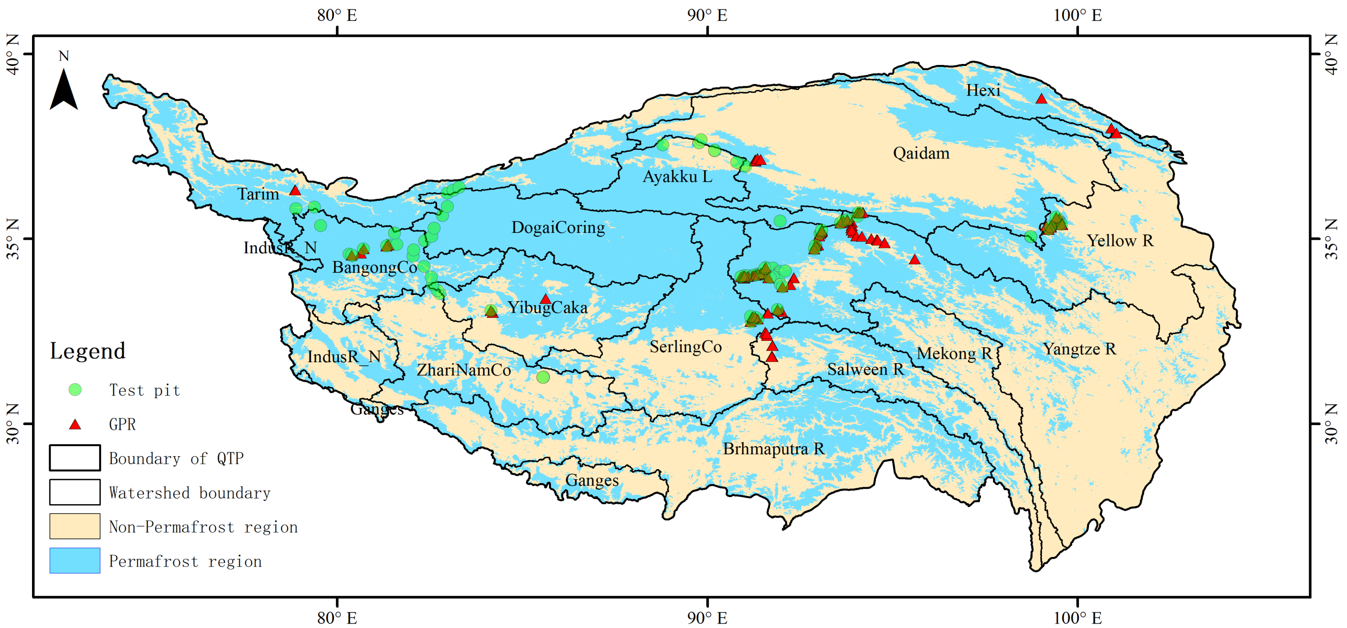

Figure 1: Study area map showing the extent of the Qinghai–Tibet Plateau permafrost region, watershed, and the spatial distribution of the 342 field-surveyed ALM sample points (circles: core-sampling-derived; triangles: GPR-derived). The permafrost region extent is derived from Zou et al. (2025), and the boundary of QTP is derived from Zhang et al. (2021)

The intrinsic uncertainty of the GPR-derived ALM was quantified through a three-method cross-validation framework at 18 calibration pits, in which wave propagation velocities were independently determined using (i) back–forward wide-angle reflection and refraction (BFWARR), (ii) multi-offset soundings (MOS), and (iii) pit profile calibration (PPC) with a metal plate buried at the frozen–thawed interface (Du et al., 2020). The comparison yielded a maximum absolute velocity error of 0.008 m ns^{-1} and an average absolute error of 0.004 m ns^{-1}; when the velocity interpretation error remains within this average level, the CRIM calibration (Eq. 2) produces a maximum ALM error of 0.08 m^3 m^{-3}. Direct validation against the 18 co-located pit measurements (observed ALM range 0.124–0.614 m^3 m^{-3}) showed that the GPR-derived ALM is virtually unbiased: bias = +0.0003 m^3 m^{-3}, MAE = 0.027 m^3 m^{-3}, RMSE = 0.032 m^3 m^{-3}, R^2 = 0.95, and Lin's concordance correlation coefficient = 0.97, with 88.9% of the paired differences within ±0.05 m^3 m^{-3} and 100% within ±0.10 m^3 m^{-3}. The representativeness of the calibration pits was confirmed against 128 additional active-layer pits across the QTP, spanning organic matter contents of 0.3–17.5% and bulk densities of 0.6–1.8 g cm^{-3}.

All GPR surveys were conducted during the peak-thaw season (September to October), when the active layer is fully thawed at its annual maximum thickness and liquid water content reaches its seasonal maximum before the onset of autumn freeze-back. This timing ensures climatologically meaningful characterisation of the ALM state and optimal reflection signal quality at the frozen–thawed interface.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

2.2 Sentinel-1 Backscatter Difference Index

To characterize spatial variations in surface wetness conditions related to ALM, a backscatter difference index (DeltaSigma, Δσ) was constructed based on Sentinel-1 Ground Range Detected (GRD) products. Interferometric Wide (IW) mode, VV-polarization data were used, and the seasonal contrast between the wet season (June) and the frozen dry season (January–February) was utilized to delineate the seasonal variation in surface and near-surface moisture status. Sentinel-1 ascending and descending orbit images from 2015 to 2024 were used. The IW-mode GRD products are distributed at a nominal pixel spacing of 10 m × 10 m, corresponding to an effective spatial resolution of approximately 20 m × 22 m in the range and azimuth directions. All images were first radiometrically calibrated, and the backscatter coefficients in the decibel (σ^0_{dB},dB) domain were converted to linear power domain (σ^0_{pow},Eq. 3). Topographic correction was then applied via incidence angle normalization and volumetric scattering normalization (Eqs. 4–5) to eliminate the effects of terrain relief and radar observation geometry. The corrected backscatter coefficients were then converted back to the dB domain (Eq. 6).

$$\sigma^0_{pow}=10^{\sigma^0_{dB}/10}, \tag{3}$$

$$\gamma^0=\sigma^0_{pow}/\cos(\theta_i), \tag{4}$$

$$\gamma^0_{vol}=\gamma^0/V, \tag{5}$$

$$\gamma^0_{vol,dB}=10\log_{10}(\gamma^0_{vol}), \tag{6}$$

where θ_i is the local incidence angle in radians, and V is the volume scattering correction factor, defined as $V=\left|\tan\left(\frac{\pi}{2}-\theta_i+\alpha_r\right)/\tan\left(\frac{\pi}{2}-\theta_i\right)\right|$, where α_r is the slope component in the radar viewing direction (computed from terrain slope and relative azimuth angle).

To further reduce the systematic effects of per-pixel incidence angle variation, the backscatter coefficients were normalized to a reference incidence angle of 38° based on statistical relationships, with separate linear normalization models established for ascending and descending orbit data (Eq. 7) (Bauer-Marschallinger et al., 2021; Li et al., 2025):

$$\sigma=\gamma^0_{vol,dB}+b(angle-38), \tag{7}$$

The minimum backscatter value during January–February of each year was used as the frozen dry season baseline, and the monthly mean backscatter of June was used to represent wet season conditions. The multi-year average annual DeltaSigma ($\Delta\sigma$) over 2015–2024 was calculated as Eq. 8. For areas within the study region lacking ascending orbit data, interpolation was performed using a linear statistical relationship between ascending and descending DeltaSigma values (Eq. 9) to improve spatial completeness:

$$\Delta\sigma = \frac{1}{10}\sum_{i=2015}^{i=2024}\left(\sigma_{mean,i} - \sigma_{min}\right) , \tag{8}$$

$$\Delta\sigma_{ASC} = 0.96 \times \Delta\sigma_{DESC} - 0.82 , \tag{9}$$

2.3 Normalized Difference Vegetation Index

Vegetation conditions were characterized using the Normalized Difference Vegetation Index (NDVI). NDVI was calculated from Landsat Collection 2 Level-2 surface reflectance products, including data from Landsat 5 TM, Landsat 7 ETM+, Landsat 8 OLI, and Landsat 9 OLI-2, all distributed at a spatial resolution of 30 m. All data acquisition and processing were conducted on the Google Earth Engine (GEE) platform, covering the period 2014–2024. Low-quality pixels affected by cloud, cloud shadow, and other contamination were masked using the QA_PIXEL band, and NDVI was calculated from the appropriate red and near-infrared bands according to each sensor. To minimize the influence of non-growing-season observations, only images acquired between Day of Year (DOY) 120 and 300 were retained. The annual maximum NDVI and median NDVI were calculated per pixel; when the date corresponding to the maximum NDVI fell outside the growing season, the median NDVI was substituted. A multi-year pixel-wise median composite was then computed across annual NDVI results from 2014 to 2024 to obtain a long-term representative NDVI dataset characterizing the maximum vegetation cover conditions over the study area at multi-year timescales.

2.4 Normalized Difference Water Index

The Normalized Difference Water Index (NDWI) was incorporated as a predictor variable for ALM spatial modeling. In this study, NDWI is used to characterize spatial variations in long-term surface and near-surface wetness conditions rather than for open water identification. In non-water-covered areas, NDWI primarily reflects changes in liquid water content within surface soil and vegetation canopies, with its physical basis derived from the different responses of near-infrared (NIR) and shortwave infrared (SWIR) reflectance to water content(Gao, 1996). NDWI data were sourced from the Landsat Collection 2 Level-2 8-day composite NDWI product (LANDSAT/COMPOSITES/C02/T1_L2_8DAY_NDWI), accessed via the GEE platform. This product is generated from atmospherically corrected Landsat surface reflectance data with a spatial resolution of 30 m. All available NDWI observations from 1 January 2014 to 31 December 2024 were selected. To highlight the long-term spatial pattern of surface wetness conditions and reduce the influence of short-term hydrological events such as precipitation and snowmelt, a pixel-wise multi-year minimum composite was computed as the static predictor for active layer moisture mapping.

2.5 Land Surface Temperature

Land Surface Temperature (LST) was incorporated as an important feature variable for characterizing the spatial distribution of ALM. There is a well-established physical linkage between LST and near-surface soil moisture: increased soil moisture enhances latent heat flux and suppresses sensible heat flux, thereby reducing surface warming and resulting in relatively lower land surface temperatures under equivalent radiative forcing conditions (Sandholt et al., 2002). LST data were sourced from the Landsat 8 Collection 2 Level-2 surface temperature product (LANDSAT/LC08/C02/T1_L2), which retrieves land surface temperatures from thermal infrared band observations at a spatial resolution of 30 m. All available imagery from 2014 to 2024 was selected. LST was calculated from the thermal infrared Band 10 (ST_B10) using the official Landsat Collection 2 scale factors:

$$LST = ST_{B10} \times 0.00341802 + 149 , \tag{10}$$

where LST is the land surface temperature in Kelvin (K). Quality control was applied using the QA_PIXEL band to exclude pixels affected by expanded cloud, cirrus, cloud shadow, and snow. Anomalously low temperature values (< 200 K) were also masked. A multi-year summer maximum LST composite was constructed based on all valid observations during June–September of 2014–2024. The maximum value composite was adopted based on two considerations: (1) it maximizes spatial data completeness by minimizing gaps caused by cloud cover, varying observation conditions, and incomplete orbit coverage; and (2) the spatial variability of maximum summer LST is partly constrained by soil moisture conditions, as higher soil water content suppresses surface temperature rise through enhanced evapotranspiration, so multi-year maximum LST can reflect the relative spatial variation in ALM status.

2.6 Topographic Factors

Ten topographic factors were derived from the SRTM90_V4 digital elevation model (DEM), including: Slope, Aspect, Profile Curvature, Plan Curvature, Topographic Contributing Area (TCA), Topographic Wetness Index (TWI), Relative Slope Position (RSP), LS Factor (slope length-gradient factor), Convergence Index (CI), and Valley Depth (VD). The DEM data were acquired from the CGIAR-CSI-processed SRTM90_V4 dataset, a void-filled 3 arc-second (~90 m) DEM product, via GEE, projected into an Albers equal-area coordinate system to ensure consistency in spatial analysis and area calculations. Two DEM pre-processing strategies were applied to ensure the physical validity of each factor: (1) a projected, non-depression-filled DEM was used for gradient and curvature variables (Slope, Aspect, Profile Curvature, Plan Curvature, RSP, CI, VD) to preserve original terrain characteristics; and (2) a hydrologically conditioned DEM (breaching-first combined with limited filling) was used for watershed and flow-related variables (TCA, TWI, LS Factor) to avoid flow direction artefacts. In the modelling framework, Aspect (a circular variable ranging 0–360°) was transformed into sin(Aspect) and cos(Aspect) to avoid discontinuity at the 0°/360° boundary. The topographic factors reflect terrain relief, slope position, and flow accumulation characteristics, providing physically meaningful information on how topography controls the spatial distribution of soil moisture through processes such as lateral drainage, solar radiation condition, and water accumulation. Detailed formulations, definitions, and the physical rationale linking each topographic factor to soil moisture are provided in the Supplementary Table S1.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

2.7 Soil Physicochemical Properties

Soil physicochemical properties derived from the China Soil Dataset for Land Surface Modeling (CSDLv2) at a spatial resolution of 90 m (Shi et al., 2025), were obtained from the National Tibetan Plateau Data Center (TPDC). This dataset was constructed using digital soil mapping approaches based on multi-source soil survey data and environmental covariates, and it well represents the spatial heterogeneity of soil physicochemical properties across China, including the QTP. CSDLv2 provides soil property information at six standard depth layers (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm), including: soil pH, particle size distribution (sand, silt, clay content), bulk density, organic carbon content, gravel content, alkali-hydrolyzable nitrogen (AN), total nitrogen (TN), cation exchange capacity (CEC), porosity, total potassium (TK), total phosphorus (TP), available potassium (AK), available phosphorus (AP), and soil color.

Considering that active layer moisture is jointly influenced by soil physical and chemical properties across shallow to mid-depth layers, Spearman correlation analyses were performed between multi-depth soil properties and field-measured ALM values. For each soil property category, the different depth layer (e.g. 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) showing the highest correlation with ALM was selected as the model input feature to avoid multicollinearity and to highlight the most influential depth intervals. The final selected soil property features include: AK_60–100, AN_5–15, AP_5–15, BD_60–100, CEC_5–15, clay_0–5, clay_100–200, gravel_30–60, OC_0–5, pH_5–15, porosity_15–30, sand_60–100, silt_0–5, TK_100–200, TN_0–5, and TP_0–5. These soil attributes collectively govern active layer moisture storage and transport processes through mechanisms including soil texture, water retention capacity, nutrient supply, and soil structure.

2.8 Climate predictor: potential evapotranspiration

To explicitly represent the climatic water–energy control on ALM, annual potential evapotranspiration (PET) was incorporated as a predictor variable. PET was obtained from the ChinaMet 0.01° (approximately 1 km) gridded dataset of annual Penman–Monteith PET (petPM; citation to be added) for 2009–2024. Annual layers were averaged pixel-wise (ignoring years with missing data) and bilinearly resampled to the common 90 m grid. The multi-year mean PET (pet_mean) characterises the long-term climatic evaporative demand which, together with moisture supply, governs the dry–wet gradient across the QTP. Because pet_mean represents a regional-scale climatic gradient that substantially exceeds the sampling footprint, it was treated as a fixed predictor and added to the models after the SHAP-RFE procedure rather than being subjected to sample-based feature elimination (Sect. 2.9.2).

The 32 predictors collectively span four groups: satellite-derived moisture, vegetation, and thermal indices (DeltaSigma, NDVI, NDWI, and LST), eleven topographic variables derived from the SRTM DEM (ten factors, with Aspect represented by its sine and cosine transforms), sixteen depth-specific soil physicochemical properties from CSDLv2, and the climatic predictor pet_mean. Table 1 summarises the definition, source dataset, native spatial resolution, and temporal compositing of all predictor variables.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Table 1: Summary of the predictor variables used for ALM modelling, listing the definition, source dataset, native spatial resolution, and temporal compositing of each variable. All predictors were resampled to the common 90 m Albers equal-area analysis grid (Sect. 2.9.1); pet_mean was bilinearly resampled from its native 0.01° grid. Detailed topographic formulations and the per-model final feature subsets are provided in Tables S1 and S2 of the Supplementary.

Predictor	Description / definition	Source dataset	Native spatial resolution
Remote sensing indices			
DeltaSigma	Seasonal backscatter difference index: multi-year (2015–2024) mean of the difference between June mean and January–February minimum VV backscatter	Sentinel-1 GRD, IW mode, VV (COPERNICUS/S1_GRD on GEE)	10 m pixel spacing (≈ 20 m × 22 m)
NDVI	Multi-year (2014–2024) median of annual growing-season (DOY 120–300) maximum NDVI	Landsat 5/7/8/9 Collection 2 Level-2 surface reflectance	30 m
NDWI	Multi-year (2014–2024) pixel-wise minimum composite of 8-day NDWI	Landsat Collection 2 Level-2 8-day NDWI composite	30 m
LST	Multi-year (2014–2024) summer (June–September) maximum land surface temperature	Landsat 8 Collection 2 Level-2 surface temperature (ST_B10)	30 m
Topographic factors			
Slope	Local slope gradient	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
Aspect_sin	Sine transform of slope aspect (0–360°)	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
Aspect_cos	Cosine transform of slope aspect (0–360°)	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
Profile curvature	Rate of change of slope gradient along the downslope profile	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
Plan curvature	Terrain curvature perpendicular to the downslope direction	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
TCA	Topographic contributing area	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
TWI	Topographic wetness index	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
RSP	Relative slope position	SRTM90_V4 DEM (CGIAR-CSI,	3 arc-seconds (~90 m)

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

		void-filled)	
LS_factor	Slope length–gradient (LS) factor	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
CI	Convergence index	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
VD	Valley depth	SRTM90_V4 DEM (CGIAR-CSI, void-filled)	3 arc-seconds (~90 m)
Soil physicochemical properties (depth layer selected by Spearman correlation; Sect. 2.7)			
AK_60–100	Available potassium, 60–100 cm depth (mg kg^{-1})	CSDLv2 (TPDC)	90 m
AN_5–15	Alkali-hydrolysable nitrogen, 5–15 cm depth (mg kg^{-1})	CSDLv2 (TPDC)	90 m
AP_5–15	Available phosphorus, 5–15 cm depth (mg kg^{-1})	CSDLv2 (TPDC)	90 m
BD_60–100	Bulk density, 60–100 cm depth (g cm^{-3})	CSDLv2 (TPDC)	90 m
CEC_5–15	Cation exchange capacity, 5–15 cm depth (me/100 g)	CSDLv2 (TPDC)	90 m
clay_0–5	Clay content, 0–5 cm depth (%)	CSDLv2 (TPDC)	90 m
clay_100–200	Clay content, 100–200 cm depth (%)	CSDLv2 (TPDC)	90 m
gravel_30–60	Rock fragment content, 30–60 cm depth (g/100 g)	CSDLv2 (TPDC)	90 m
OC_0–5	Soil organic carbon, 0–5 cm depth (g/100 g)	CSDLv2 (TPDC)	90 m
pH_5–15	Soil pH (H_2O), 5–15 cm depth (–)	CSDLv2 (TPDC)	90 m
porosity_15–30	Soil porosity, 15–30 cm depth (cm^3 cm^{-3})	CSDLv2 (TPDC)	90 m
sand_60–100	Sand content, 60–100 cm depth (%)	CSDLv2 (TPDC)	90 m
silt_0–5	Silt content, 0–5 cm depth (%)	CSDLv2 (TPDC)	90 m
TK_100–200	Total potassium, 100–200 cm depth (g/100 g)	CSDLv2 (TPDC)	90 m
TN_0–5	Total nitrogen, 0–5 cm depth (g/100 g)	CSDLv2 (TPDC)	90 m
TP_0–5	Total phosphorus, 0–5 cm depth (g/100 g)	CSDLv2 (TPDC)	90 m
Climate			
pet_mean	Multi-year (2009–2024) mean annual Penman–Monteith potential	ChinaMet petPM	0.01° (≈ 1 km)

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

	evapotranspiration

2.9 Machine Learning Modelling

2.9.1 Model Selection and Ensemble Framework

Four ensemble tree-based machine learning models were used. Random Forest reduces prediction variance via bootstrap aggregating and random feature sub-sampling, offering robustness for multi-source variable prediction (Breiman, 2001). Extra Trees employs fully randomized split thresholds to further reduce sensitivity to local sample distributions and to mitigate overfitting under complex terrain conditions (Geurts et al., 2006). XGBoost incorporates regularization terms within a gradient boosting framework, enabling efficient learning of nonlinear responses to temperature, vegetation, and microwave signals (Chen and Guestrin, 2016). CatBoost employs an ordered boosting strategy to reduce prediction bias, maintaining high stability even with limited sample sizes and high-dimensional features (Dorogush et al., 2018).Collectively, these four algorithms span the two dominant ensemble paradigms—bagging (Random Forest, Extra Trees), which primarily reduces prediction variance, and gradient boosting (XGBoost, CatBoost), which primarily reduces prediction bias—so that their prediction errors arise from structurally different learning mechanisms and are not fully correlated, a prerequisite for effective multi-model fusion (Sect. 2.9.3).

All four models were trained on the same sample dataset ($n = 342$) using a unified initial feature set of 32 variables, comprising: remote sensing indices (DeltaSigma, NDVI, NDWI, LST), a climate predictor (multi-year mean annual PET; Sect. 2.8), soil physicochemical properties (16 variables), and topographic factors (including sin(Aspect) and cos(Aspect) as transformations of the original Aspect variable). All covariates (Table 1) were resampled to a common 90 m Albers equal-area grid to ensure spatial consistency. Of these, the 31 non-climatic predictors were subjected to SHAP-RFE, and pet_mean was subsequently added as a fixed predictor to the selected subsets (Sect. 2.9.2).

2.9.2 Model Training, Feature Selection, and Accuracy Evaluation

To prevent information leakage during simultaneous model tuning and feature selection, a nested cross-validation (Nested CV) framework was adopted (Cawley and Talbot, 2010; Varma and Simon, 2006). The outer loop employed 5-fold cross-validation to evaluate generalization performance, while the inner loop used 4-fold cross-validation for hyper parameter optimization.

SHAP-based Recursive Feature Elimination (SHAP-RFE) was integrated to identify parsimonious, physically interpretable feature subsets (Lundberg and Lee, 2017). In each SHAP-RFE iteration, SHAP feature importance values were computed, and the feature with the lowest mean absolute SHAP contribution was removed. The complete Nested CV procedure was then repeated to evaluate the impact of feature reduction on model performance stability, generating a performance curve as a function of feature count. The feature subset that maximized the mean outer-fold adjusted R^2 (Adj R^2) was designated as

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

the optimal combination for each model. The hyper parameter configuration from the best-performing inner fold was used to retrain the final model on the complete dataset for spatial prediction. Because pet_mean characterises the climatic water–energy balance at a regional scale far exceeding the sampling footprint, it was excluded from the SHAP-RFE procedure and subsequently added as a fixed predictor to the optimal subset of each model, so as to explicitly represent the regional climatic gradient and strengthen cross-regional transferability.

Model performance was evaluated using four metrics computed across outer cross-validation folds: coefficient of determination (R^2, Eq. 11), adjusted coefficient of determination (Adj R^2, Eq. 12), root mean squared error (RMSE, Eq. 13), and mean absolute error (MAE, Eq. 14), with mean and standard deviation statistics computed across different folds to quantify model prediction performance and uncertainty.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}_i \right)^2}, \tag{11}$$

$$Adj\,R^2 = 1 - \left(1 - R^2 \right) \frac{n-1}{n-p-1}, \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n} \sum \left(y_i - \hat{y}_i \right)^2}, \tag{13}$$

$$MAE = \frac{1}{n} \sum \left| y_i - \hat{y}_i \right|, \tag{14}$$

where y_i、 $\hat{y}_i$ and $\overline{y}_i$ represent observed values, predicted values, and the mean of observed values, respectively, n is the sample size, and p is the number of features.

Random k-fold cross-validation can yield over-optimistic performance estimates when training and test observations are spatially autocorrelated, because random partitioning leaves observations from every region in each fold and allows the model to exploit spatial proximity (Ploton et al., 2020). Because the intended application of this dataset is to predict ALM in areas without field observations, we additionally implemented a group-based spatial cross-validation scheme that explicitly tests model transferability to unsampled regions (Meyer and Pebesma, 2021).

The 342 observations were grouped according to their 12 major watersheds (Table 4) into four spatially contiguous folds: G1_East (Yangtze R., n = 158), G2_Northeast (Yellow R. + Qaidam, n = 86), G3_CentralLakes (six interior endorheic lake basins, n = 75), and G4_Peripheral (Tarim + Hexi + Salween R., n = 23). In each iteration, the entire group was withheld as

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

the test set while the model was trained on the remaining samples, mimicking the operational scenario of predicting ALM in an entirely unobserved region. An alternative five-fold grouping and a strict leave-one-basin-out (LOBO) scheme were additionally implemented to assess the sensitivity of the results to the partition definition. As argued by Wadoux et al. (2021), spatial cross-validation deliberately validates the sub-areas of the map that are environmentally most distant from the calibration points, and therefore provides a conservative lower bound of model transferability rather than an estimate of map accuracy per se (see Sect. 4.4). The corresponding spatial cross-validation results are presented in Sect. 3.2.

2.9.3 Bias-Aware Multi-Model Fusion Strategy

To correct the directional biases of the individual models diagnosed in Sect. 3.4, a bias-aware multi-model fusion strategy was developed.Multi-model fusion was adopted instead of selecting a single best model for two reasons. First, the four models were statistically indistinguishable on the spatial-CV hold-out predictions (pooled R^2 = 0.296–0.382; Sect. 3.2), so identifying a single optimal model would be neither defensible nor robust to the cross-validation design. Second, each architecture retained a partially distinct SHAP-RFE predictor subset (Sect. 2.9.2) and embodies a different bias–variance trade-off, so the four prediction surfaces carry complementary information: fusion suppresses model-specific idiosyncratic errors while exploiting the systematically opposing bias tendencies of the bagging and boosting families at the ALM distribution tails—biases that no individual model corrects on its own (Sect. 3.4). Based on the spatial predictions from all four models, three ensemble statistics were computed per pixel: minimum, mean, and maximum. The fusion rule was then applied as follows:

(i) For each pixel, the minimum, mean, and maximum of the four model predictions were computed. The fused value was then obtained by linearly blending these three ensemble statistics as a function of the multi-model mean ($mean_4$): (i) when $mean_4$ is below 0.2 $m^3\ m^{-3}$, the minimum prediction was increasingly weighted to counteract overestimation in the low-ALM tail, reaching full weight at $mean_4 \le 0.15\ m^3\ m^{-3}$; (ii) when $mean_4$ is above 0.4 $m^3\ m^{-3}$, the maximum prediction was increasingly weighted to counteract underestimation in the high-ALM tail, reaching full weight at $mean_4 \ge 0.45\ m^3\ m^{-3}$; (iii) in the intermediate core range (0.2–0.4 $m^3\ m^{-3}$), the pixel-wise mean was retained as the consensus estimate. A ±0.05 $m^3\ m^{-3}$ linear transition window was imposed around each threshold to eliminate structural histogram discontinuities while preserving the physically motivated bias correction. This fusion approach explicitly exploits the complementary bias characteristics of different model architectures—Boosting models (CatBoost, XGBoost) tend to produce more extreme dry predictions, while Bagging models (Random Forest, Extra Trees) produce more moderate estimates—to reduce systematic errors in a physically motivated manner.

The 0.2 and 0.4 $m^3\ m^{-3}$ thresholds were determined from the residual structure diagnosed during the spatial cross-validation (Sect. 3.2). To evaluate the sensitivity of the final product to the fusion rule, the bias-aware strategy was compared against four alternatives (simple mean, R^2-weighted mean, inverse-RMSE-weighted mean, and median) on the spatial-CV hold-out

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

predictions (Table 3). The fusion rule therefore contains no fitted parameters: both thresholds are read directly from the sign transitions of the spatial-CV residuals (Sect. 3.2), making the rule fully reproducible and free of arbitrary tuning.

2.9.4 Training data representativeness and area of applicability

The representativeness of the 342 training samples relative to the QTP permafrost domain was quantified following the framework of Meyer and Pebesma (2021). For each model, the dissimilarity index (DI) was computed in the multidimensional predictor space defined by the model's final feature subset (Table S2), with predictors weighted by their relative importance extracted from the trained model. For each pixel of the permafrost domain, DI was calculated as the minimum Euclidean distance to the nearest training sample in the weighted, standardised predictor space, normalised by the mean pairwise distance among training samples. The area of applicability (AOA) was derived by thresholding DI at the outlier-removed maximum of the training-data DI distribution (upper whisker, Q3 + 1.5 × IQR), computed within the spatial fold structure to remain consistent with the spatial validation (Sect. 2.9.2). The final, fused AOA was defined as the intersection of the four model-specific AOA masks, representing the region where all four models are jointly applicable.

All coverage statistics are computed over the permafrost zone only, using the 90 m permafrost distribution raster in which pixel value 1 denotes permafrost. All 342 training samples are located within, or immediately adjacent to, mapped permafrost, and field verification confirmed permafrost occurrence at all sampling locations, so no samples were excluded owing to co-registration offsets between sample coordinates and the permafrost raster. Non-permafrost pixels within the study boundary are set to no-data in all delivered layers (ALM product, prediction SD, DI, AOA, and per-pixel expected RMSE).

3 Results

3.1 Correlations Between ALM and Environmental Covariates

Spearman rank correlation analysis revealed significant and structured associations between ALM (showed as VWC in the Fig) and both terrain–remote sensing factors and soil physicochemical properties (Fig. 2).

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

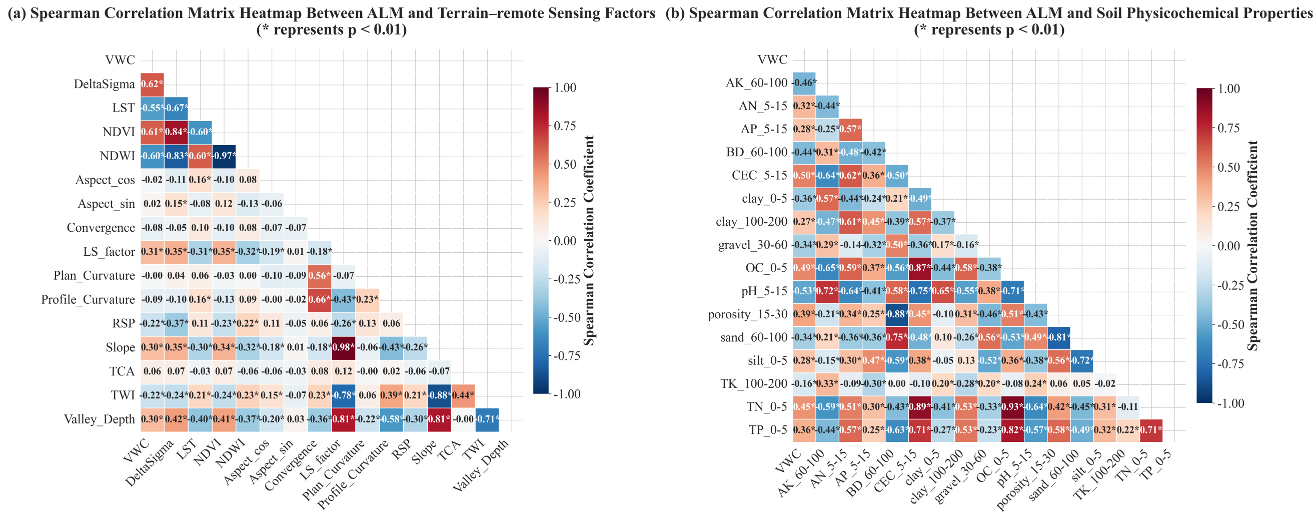

Figure 2: Spearman correlation matrices between ALM and (a) terrain–remote sensing factors and (b) soil physicochemical properties. Significant correlations (p < 0.01) are marked with asterisks.

In the terrain–remote sensing factor correlation matrix (Fig. 2a), ALM exhibited strong positive correlations with vegetation and surface moisture-related indices, particularly DeltaSigma (r = 0.62) and NDVI (r = 0.61), while showing significant negative correlations with NDWI (r = −0.60) and LST (r = −0.55). These relationships indicate that active layer moisture is tightly coupled with surface hydrothermal status and vegetation activity. Additionally, ALM was significantly correlated with topographic and hydrological parameters including the LS Factor, Slope, and Valley Depth (|r| ≈ 0.3), reflecting the control of terrain on surface water convergence and redistribution. The terrain–remote sensing factors exhibited pronounced inter-variable clustering: vegetation and moisture indices (NDVI, NDWI, DeltaSigma, LST) showed very high mutual correlations (|r| up to 0.97), and topographic parameters (Slope, LS Factor, TWI, Valley Depth) exhibited high internal correlations (|r| up to 0.98), indicating strong multicollinearity that motivated the use of machine learning models combined with SHAP-RFE for feature selection.

The soil physicochemical property correlation matrix (Fig. 2b) revealed widespread significant associations between ALM and subsurface soil attributes. ALM showed significantly positive correlations with CEC (r = 0.50), OC (r = 0.49), TN (r = 0.45), porosity (r = 0.39), and TP (r = 0.36), and significantly negative correlations with soil pH (r = −0.53), bulk density (r = −0.44), sand content (r = −0.34), and gravel content (r = −0.34). These results indicate that fine-textured, structurally loose, and organically enriched soils retain more ALM, while coarse-textured and compacted soils restrict water storage capacity.

3.2 Model Performance and Optimal Feature Subsets

Figure 3 illustrates the dynamic response of mean R^2 and mean Adj R^2 to varying numbers of retained features during the SHAP-RFE process. All four models showed a consistent pattern: as the number of features decreased from 31 to 1, mean R^2 remained relatively stable near 0.6, while mean Adj R^2 exhibited a clear initial increase—reflecting the elimination of

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

redundant variables—before stabilising and converging towards the mean R^2 value. This indicates that including the full feature set introduced substantial information redundancy without improving generalisation performance.

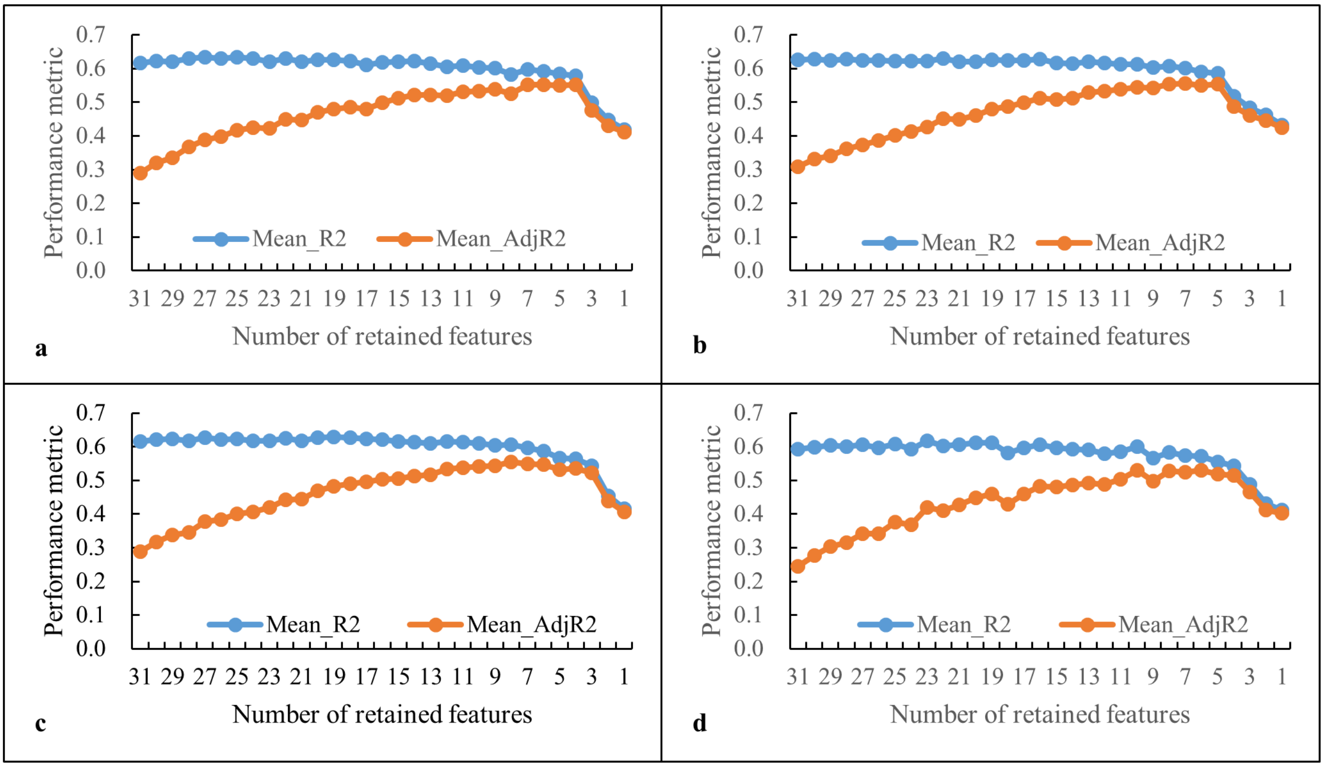

Figure 3: SHAP-RFE performance curves showing mean R^2 and mean Adj R^2 as a function of the number of retained features for (a) CatBoost, (b) Extra Trees, (c) Random Forest, and (d) XGBoost.

Using maximum mean Adj R^2 as the selection criterion, CatBoost and Extra Trees both achieved peak performance with 7 retained features. CatBoost selected: DeltaSigma, LST, NDVI, NDWI, AP_5–15, OC_0–5, and LS_factor. Extra Trees retained: DeltaSigma, LST, NDVI, NDWI, OC_0–5, CEC_5–15, and pH_5–15, indicating greater sensitivity to soil chemical properties. Random Forest achieved optimal performance with 8 features: DeltaSigma, LST, NDVI, NDWI, AK_60–100, CEC_5–15, pH_5–15, and porosity_15–30. XGBoost required 10 features: DeltaSigma, LST, NDVI, AK_60–100, AP_5–15, OC_0–5, porosity_15–30, Aspect_cos, LS_factor, and RSP, reflecting its ability to integrate multi-dimensional terrain–soil–moisture interactions. Notably, XGBoost did not retain NDWI, likely because its information content was substituted by the highly correlated DeltaSigma or NDVI. To explicitly represent the regional climatic gradient, the multi-year mean PET (pet_mean) was subsequently added as a fixed predictor to the SHAP-RFE-selected subset of each model (Sect. 2.9.2), yielding final feature sets of 8, 8, 9, and 11 variables for CatBoost, Extra Trees, Random Forest, and XGBoost, respectively.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

Figure 4 presents SHAP-based feature importance rankings for each model using its final feature subset. Remote sensing factors associated with surface moisture processes (DeltaSigma, NDVI, NDWI, LST) consistently dominated predictions across all four models. In CatBoost (Fig. 4a), LST, DeltaSigma, and NDVI ranked in the top three, with mean SHAP contributions of 20.2%, 16.7%, and 13.4%, respectively, followed by the PET predictor pet_mean (11.0%). In Extra Trees (Fig. 4b), DeltaSigma, NDVI, and NDWI led with contributions of 23.8%, 23.2%, and 11.9%, followed by pet_mean (9.6%). In Random Forest (Fig. 4c), the top four were NDVI, DeltaSigma, LST, and pet_mean (29.7%, 26.0%, 9.1%, 7.8%). In XGBoost (Fig. 4d), NDVI, DeltaSigma, and LST contributed 30.7%, 16.0%, and 11.7%, respectively, with pet_mean ranking fifth (6.7%).

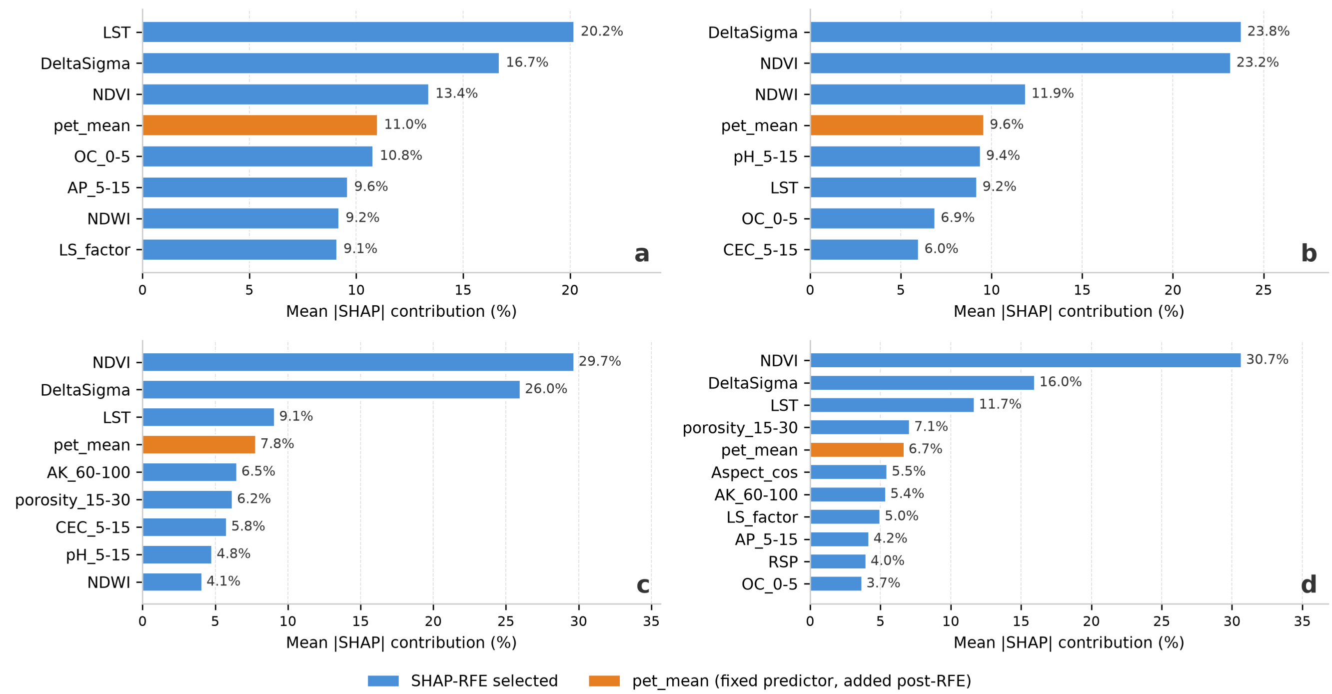

Figure 4: SHAP-based feature importance (mean absolute SHAP contribution, %) for the final models and their contributions to ALM prediction: (a) CatBoost, (b) Extra Trees, (c) Random Forest, (d) XGBoost. Blue bars denote predictors selected by SHAP-RFE; orange bars denote pet_mean, the multi-year mean potential evapotranspiration added as a fixed predictor after feature selection (Sect. 2.9.2). Final feature sets contain 8, 8, 9, and 11 predictors for CatBoost, Extra Trees, Random Forest, and XGBoost, respectively.

Soil organic carbon (OC_0–5) remained the most important non-remote-sensing variable in the three models that retained it (SHAP contributions: 10.8% in CatBoost, 6.9% in Extra Trees, and 3.7% in XGBoost), followed by soil pH, CEC, porosity, and AK at varying rankings. The multi-year mean PET (pet_mean) contributed 7–11% across all four models, ranking fourth in CatBoost (11.0%), Extra Trees (9.6%), and Random Forest (7.8%) and confirming the overarching control of climatic evaporative demand on ALM. The aggregate contribution of remote sensing factors was 58–69% across models; the climate predictor contributed 7–11%; soil physicochemical properties contributed 20–23%; and topographic factors contributed 0–15%, with XGBoost showing the highest topographic contribution (14.5%).

Table 2 summarises the cross-validation performance metrics. All four models achieved comparable accuracy: mean cross-validated R^2 = 0.62–0.63, Adj R^2 = 0.61–0.62, RMSE ≈ 0.08 $m^3\ m^{-3}$, and MAE ≈ 0.06 $m^3\ m^{-3}$. All four models retained the land-surface predictors DeltaSigma, LST, NDVI, and pet_mean, supplemented by model-specific soil (AK_60–100, CEC_5–15, pH_5–15, porosity_15–30, OC_0–5, AP_5–15), topographic (LS_factor, Aspect_cos), and moisture-related (NDWI, RSP) attributes. The final subsets comprised 8 predictors for CatBoost and Extra Trees, 9 for Random Forest, and 11 for XGBoost; the complete subsets retained by each model are listed in Table S2 of the Supplementary.

Table 2:Cross-validation performance metrics of the four models under random and group-based spatial cross-validation schemes (RMSE and MAE in $m^3\ m^{-3}$). The spatial-CV results represent conservative transferability estimates rather than map accuracy.

Model	Strategy	R^2_pooled	AdjR^2_pooled	RMSE_pooled	MAE_pooled
Random Forest	Group4	0.296	0.277	0.11	0.087
CatBoost		0.379	0.364	0.103	0.081
XGBoost		0.382	0.362	0.103	0.08
Extra Trees		0.362	0.347	0.104	0.081
Random Forest	Group5	0.3	0.282	0.109	0.087
CatBoost		0.377	0.362	0.103	0.081
XGBoost		0.373	0.352	0.103	0.08
Extra Trees		0.363	0.348	0.104	0.081
Random Forest	LOBO	0.396	0.38	0.101	0.08
CatBoost		0.424	0.41	0.099	0.078
XGBoost		0.413	0.393	0.1	0.078
Extra Trees		0.433	0.419	0.098	0.077
Random Forest	Random 5-fold	0.622	0.612	0.08	0.062
CatBoost		0.627	0.618	0.08	0.061
XGBoost		0.624	0.611	0.08	0.062
Extra Trees		0.632	0.623	0.079	0.061

Figure 5 presents observed-versus-predicted scatter plots combining all training data (Fig. 5a) and all test data (Fig. 5b) pooled across the five outer folds of spatial cross-validation. Individual models are distinguished by colour (CatBoost, Extra Trees, Random Forest, and XGBoost). Per-model counterparts, pooling all training and test data across the five outer folds with points colour-coded by spatial cross-validation fold, are provided in Fig. S2 of the Supplementary. For the training set (Fig. 5a), the pooled predictions show strong agreement with observations (R^2 = 0.93, RMSE = 0.033 $m^3\ m^{-3}$, MAE = 0.025 $m^3\ m^{-3}$, bias = +0.0002 $m^3\ m^{-3}$, slope = 0.87), indicating high in-sample fidelity. For the test set (Fig. 5b), the pooled out-of-fold predictions exhibit substantially greater scatter (R^2 = 0.36, RMSE = 0.105 $m^3\ m^{-3}$, MAE = 0.082 $m^3\ m^{-3}$, bias = −0.0083

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

m³ m⁻³, slope = 0.40), reflecting the difficulty of spatial extrapolation beyond the calibration domain. A consistent systematic bias pattern is apparent in both panels: overestimation at low ALM (< 0.2 m³ m⁻³) and underestimation at high ALM (> 0.4 m³ m⁻³), a characteristic consequence of sample distribution imbalance and limited extreme-value training data.

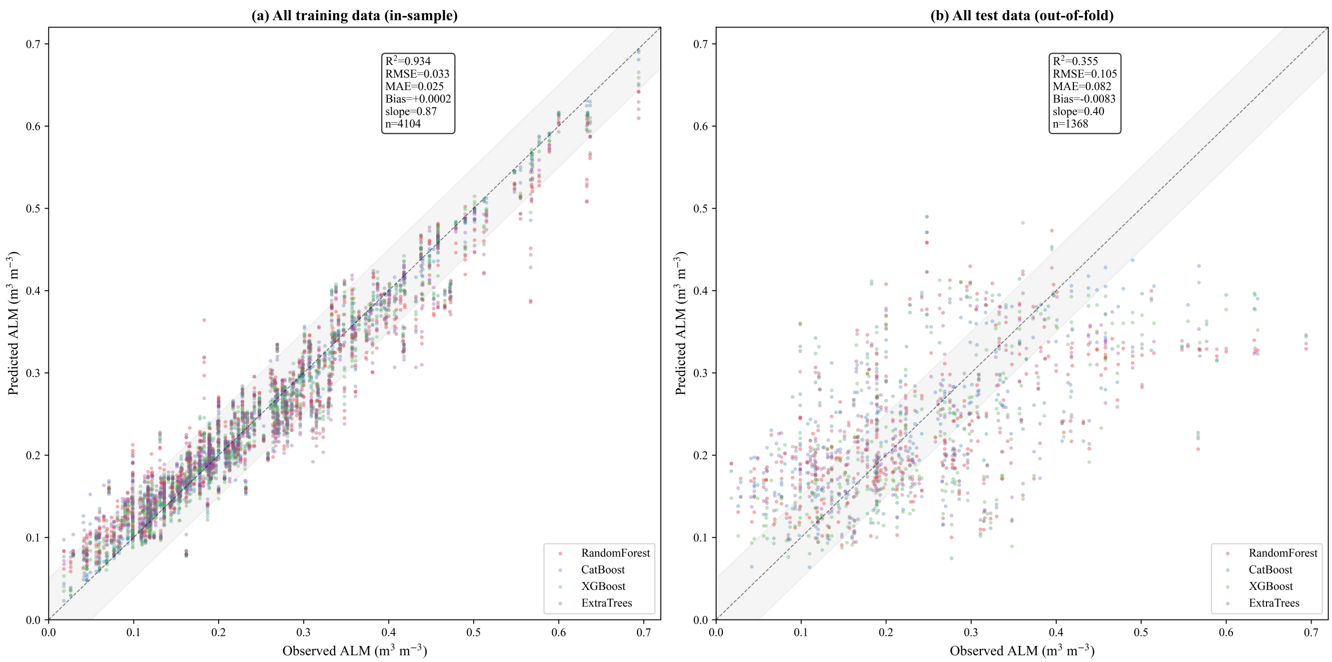

Figure 5: Observed versus predicted ALM scatter plots combining all (a) training (in-sample) and (b) test (out-of-fold) predictions pooled across the five outer folds of spatial cross-validation. Each model is represented by a distinct colour (CatBoost, Extra Trees, Random Forest, XGBoost). The dashed line indicates the 1:1 reference; the solid line shows the linear regression fit through all points.

Under the group-based spatial cross-validation, the pooled R^2 dropped to 0.30–0.38 across the four models (Table 2), reflecting the difficulty of transferring the models to hydrologically and climatically distinct basins; the five-fold spatial grouping yielded nearly identical results (0.30–0.38), and the strict 12-basin LOBO scheme yielded a higher pooled R^2 (0.40–0.43) because each fold's training set is larger. The G1_East fold (Yangtze R.) was the most demanding, with fold R^2 values ranging from −0.11 to 0.08, reflecting the high soil-moisture variability of the eastern permafrost margins. Following Wadoux et al.(2021), these spatial-CV metrics are interpreted as a conservative lower bound of model transferability rather than map accuracy per se (Sect. 4.4).

3.3 Spatial Distribution of Individual Model Predictions

Figure 6 presents the individual spatial distribution maps of ALM from each of the four models within the permafrost region. All four models show broadly consistent spatial patterns: the highest ALM values are concentrated in the Yellow River (Yellow R.) basin and adjacent areas of the eastern QTP, as well as the south eastern river basins (Yangtze R., Mekong R., Salween R.) and the eastern portion of the Brahmaputra–Yarlung Tsangpo River basin (Brahmaputra R.). The lowest ALM

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

regions are found in the Qaidam Basin and the endorheic lake basins of the Qiangtang Plateau (including Serling Co, Zhari Nam Co, Yibug Caka, Dogai Coring, and Bangong Co). The southern rim of the QTP also shows notably elevated moisture values. This broad spatial pattern is consistent with existing surface soil moisture remote sensing products for the QTP (Han et al., 2023; Li et al., 2025; Xing et al., 2025), supporting the credibility of the simulation results and confirming that surface hydrothermal conditions and their associated remote sensing indicators (DeltaSigma, NDVI, LST, NDWI) are the primary controls on ALM spatial patterns.

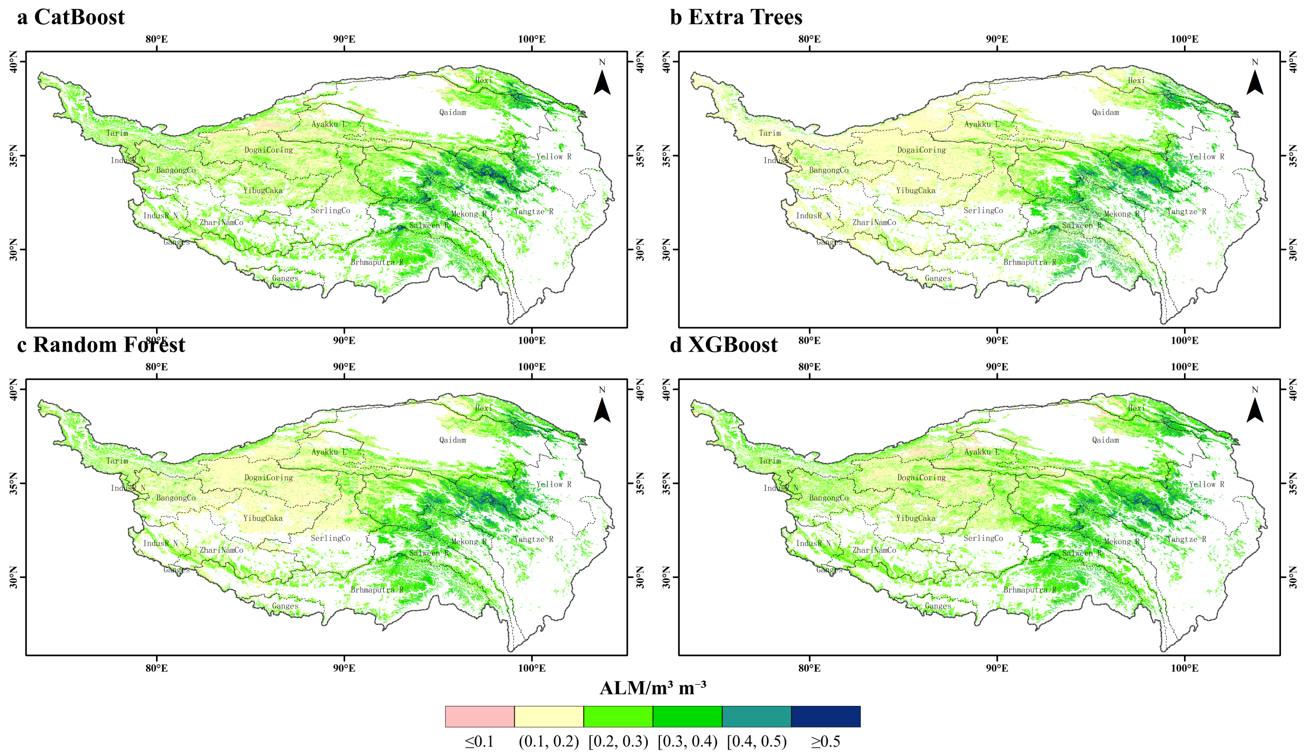

Figure 6: Spatial distribution of ALM simulated by (a) CatBoost, (b) Extra Trees, (c) Random Forest, and (d) XGBoost. Major watershed boundaries are overlaid for spatial reference.

Despite comparable overall accuracy, the four models exhibit notable differences in spatial detail and in the modelled range of extreme ALM values. CatBoost and XGBoost (Figs. 6a, 6d) produce more extreme dry conditions in the Qaidam Basin (ALM ≤ 0.1 m³ m⁻³), while Extra Trees and Random Forest (Figs. 6b, 6c) tend to yield more moderate values (0.1–0.2 m³ m⁻³) in these areas, reflecting fundamental differences in variance control between Bagging (Random Forest and Extra Trees) and Boosting (XGBoost, CatBoost) frameworks.

Figure 7 summarises the pixel proportion statistics across ALM intervals for each model. All four models assign the largest pixel fractions to the low-to-moderate moisture range (0.1–0.3 m³ m⁻³), but internal distributions differ substantially. Extra Trees concentrates the highest proportion of pixels in the (0.1, 0.2) interval (65.5%), followed by Random Forest (52.0%),

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

whereas CatBoost (41.2%) and XGBoost (39.7%) show lower concentrations in this range but markedly higher proportions in the [0.2, 0.3) interval (43.8% and 43.0%, respectively, versus 18.5% for Extra Trees and 32.2% for Random Forest). This indicates that Boosting-based models generate a broader and more evenly distributed moisture spectrum, while Bagging-based models compress predictions toward the lower moisture range. In the high-moisture range (≥ 0.3 m³ m⁻³), CatBoost yields the lowest cumulative proportion (13.4%), compared with XGBoost (15.1%), Random Forest (15.6%), and Extra Trees (15.7%), suggesting that CatBoost most strongly suppresses extreme high-value predictions. At the dry extreme (ALM ≤ 0.1 m³ m⁻³), XGBoost (2.2%) and CatBoost (1.6%) assign notably higher pixel fractions than Extra Trees (0.4%) and Random Forest (0.3%), consistent with the spatial patterns in Fig. 6.

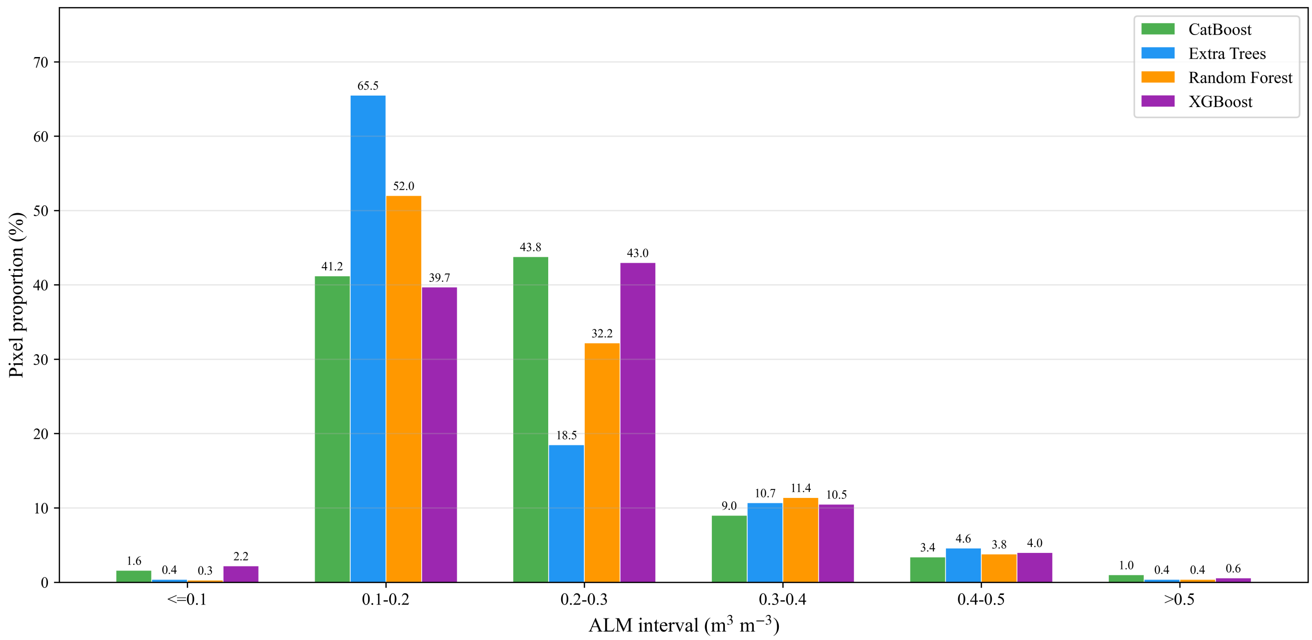

Figure 7: Pixel proportion statistics across ALM intervals for the four individual machine learning models.

3.4 Bias Characterisation and Multi-Model Fusion

Figure 8 presents the statistical relationship between field-observed mean ALM and model-simulated mean ALM aggregated at 0.1 m³ m⁻³ intervals, ordered by ascending observed ALM. All four models show near-perfect linear fits ($R^2 = 1$) at this level of aggregation, confirming that the models reliably capture the spatial trend of ALM at the plateau scale. However, all four models exhibit consistent systematic biases: overestimation at low ALM (< 0.2 m³ m⁻³) and underestimation at high ALM (> 0.4 m³ m⁻³). Extra Trees and XGBoost show near-unity regression slopes (0.98), CatBoost shows a smaller systematic bias (regression slope ≈ 0.94), while Random Forest exhibits a slightly larger deviation (slope of 0.87).

The bias-aware multi-model fusion strategy described in Sect. 2.9.3 was applied to all four model outputs. The fused product demonstrates markedly improved bias characteristics: the regression slope of observed-versus-simulated mean ALM

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

increases to0.99(Fig. 8e), substantially closer to the 1:1 reference line than any individual model, representing a significant improvement in absolute accuracy and systematic bias reduction.

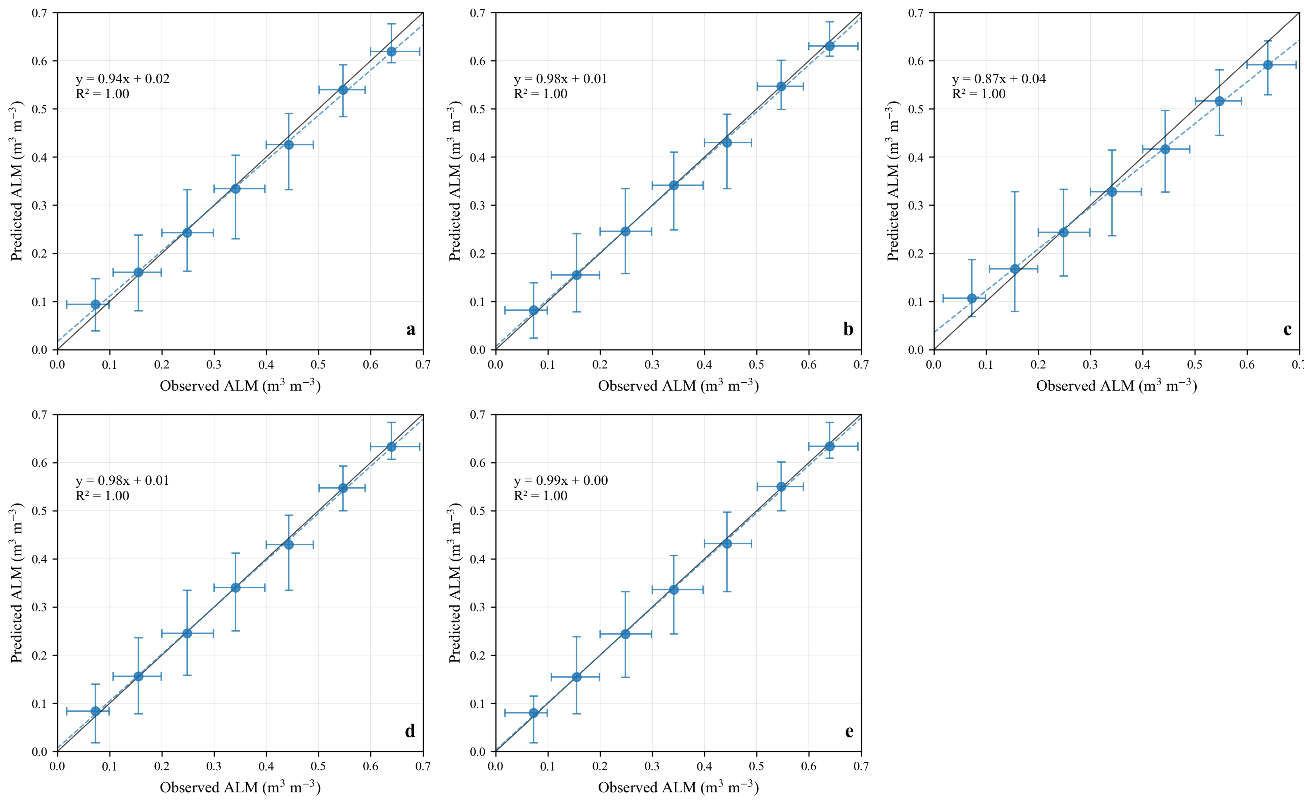

Figure 8: Relationship between observed mean ALM and simulated mean ALM aggregated at 0.1 m³ m⁻³ gradient intervals for (a) CatBoost, (b) Extra Trees, (c) Random Forest, (d) XGBoost, and (e) the fused product. Error bars denote the min–max ranges of observed (horizontal) and simulated (vertical) ALM within each 0.1 m³ m⁻³ interval.

Table 3:Sensitivity of pooled spatial-CV performance to the fusion strategy (RMSE, MAE, and bias in m³ m⁻³). The bias-aware rule is the formal scheme; alternative rules are reported to demonstrate insensitivity.

Fusion rule	Pooled R^2	Pooled RMSE	Pooled MAE	Pooled Bias
Bias-aware	0.368	0.104	0.081	−0.015
Simple mean	0.383	0.103	0.081	−0.008

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

R^2-weighted mean	0.387	0.102	0.080	−0.008
Inverse-RMSE-weighted mean	0.384	0.102	0.080	−0.008
Median	0.373	0.103	0.081	−0.008

Table 3 compares the bias-aware strategy with four alternative fusion rules on the spatial-CV hold-out predictions. All five rules yielded pooled R^2 values within a narrow range (0.368–0.387; $\Delta R^2 < 0.020$), demonstrating that the final product is largely insensitive to the choice of fusion rule; the bias-aware strategy was retained as the formal scheme for three reasons: (i) the metric differences among the five rules ($\Delta R^2 < 0.025$, $\Delta RMSE \leq 0.002$ m³ m⁻³) are an order of magnitude smaller than the retrieval uncertainty of the GPR reference data themselves (RMSE = 0.032 $m^3\ m^{-3}$; Sect. 2.1), so the nominal ranking of the rules carries no practical significance; (ii) the mean-type alternatives inherit the systematic underestimation of high-ALM pixels diagnosed in Sect. 3.4 (bin-wise bias of approximately −0.14 $m^3\ m^{-3}$ for ALM > 0.4 $m^3\ m^{-3}$) and would therefore compress the spatial value range of the product, whereas the bias-aware rule preserves the full predicted domain; and (iii) only the bias-aware rule provides a physically interpretable treatment of the diagnosed symmetric bias structure.

3.5 Final Fused ALM Dataset

Figure 9 presents the final fused ALM product integrating all four model outputs via the bias-aware fusion strategy. Because active-layer moisture (ALM) is physically meaningful only within the permafrost region, the delivered product is clipped to the QTP permafrost boundary.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

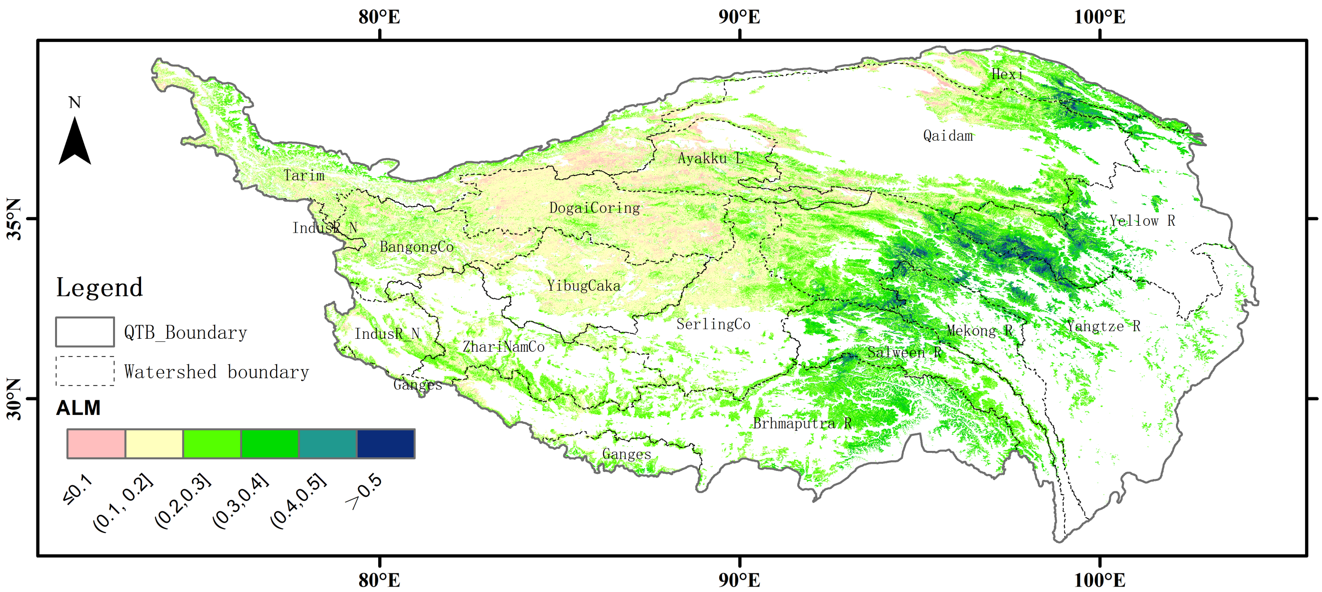

Figure 9: Final fused ALM spatial distribution across the QTP permafrost region.

Within the QTP permafrost region, the fused ALM ranges from 0.02 to 0.70 m³ m⁻³ (mean = 0.21 m³ m⁻³, SD = 0.09 m³ m⁻³). Pixel proportion statistics across ALM intervals for the permafrost zone are shown in Fig. 10. The 0.1–0.3 m³ m⁻³ range accounts for 81.6% of total pixels (47.7% in the 0.1–0.2 interval and 33.9% in the 0.2–0.3 interval), indicating that the QTP permafrost region is predominantly characterized by relatively dry active layer conditions. This dry region is concentrated in the Qiangtang Plateau endorheic lake basins (Serling Co, Yibug Caka, Dogai Coring, Bangong Co) and adjacent high-mountain areas—zones of continuous permafrost occurrence in the interior plateau where extreme cold-arid conditions simultaneously favors permafrost preservation and limit active layer moisture. High-ALM areas (0.4–0.5 m³ m⁻³: 2.7% of pixels; > 0.5 m³ m⁻³: 1.3% of pixels) are concentrated in the humid eastern and southeastern QTP.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

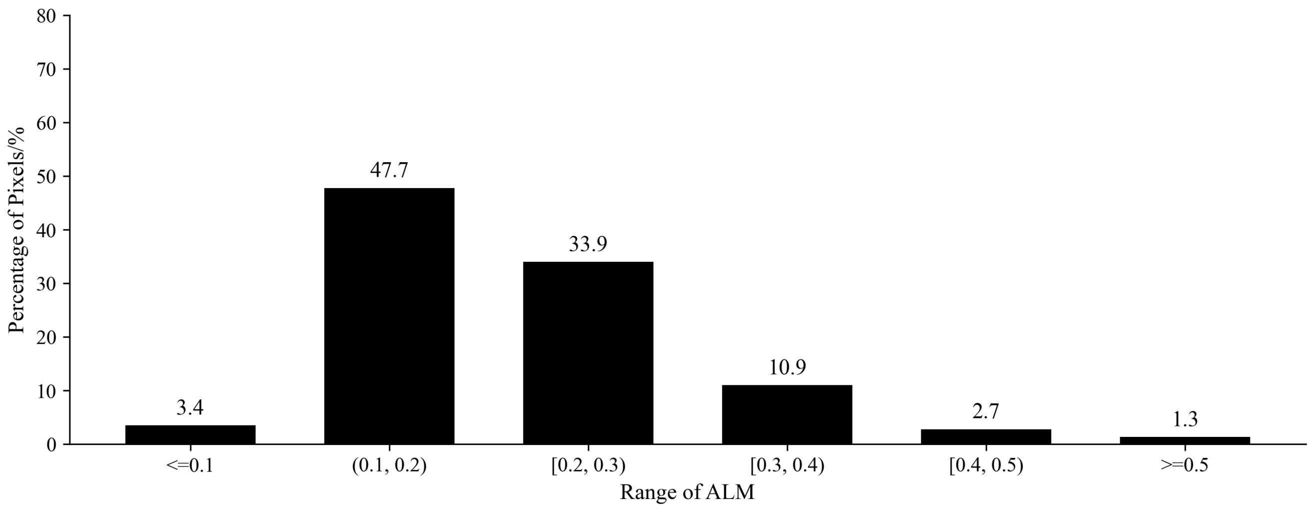

Figure 10: Pixel proportion statistics across ALM intervals for the final fused ALM in the QTP permafrost region.

Watershed-level statistics (Table 4) confirm these patterns. The watershed of Dogai Coring Lake has the lowest mean ALM (0.15 $m^3\ m^{-3}$, SD = 0.03), while the Yellow River basin has the highest (0.34 $m^3\ m^{-3}$, SD = 0.12). Basins in the central and western Qiangtang Plateau (YibugCaka, BangongCo, Tarim, Serling Co, Zhari Nam Co) all have mean ALM below the permafrost region average (0.20 $m^3\ m^{-3}$), ranging from 0.15 to 0.18 $m^3\ m^{-3}$ with low spatial variability (SD ≤ 0.05 $m^3\ m^{-3}$). Eastern and south eastern basins (Hexi, Yangtze R., Salween R., Mekong R.) exhibit higher mean ALM (0.22–0.31 $m^3\ m^{-3}$) with larger spatial variability (SD ≈ 0.10 $m^3\ m^{-3}$). The QTP permafrost region ALM thus displays a distinct southeast-high, northwest-low gradient. Furthermore, watersheds with higher mean ALM consistently exhibit greater spatial variability (higher SD), while basins with lower mean ALM show more homogeneous moisture distributions.

Table 4: ALM range, mean, and standard deviation statistics for major watersheds within the QTP permafrost region.

Name	MIN	MAX	MEAN	STD
DogaiCoring L.	0.03	0.52	0.15	0.03
YibugCaka	0.04	0.47	0.15	0.03
Ayakku L.	0.03	0.61	0.16	0.05
BangongCo	0.04	0.48	0.16	0.03
Tarim	0.03	0.58	0.17	0.04
Serling Co	0.04	0.63	0.17	0.04
ZhariNamCo	0.05	0.48	0.18	0.03
Qaidam	0.02	0.67	0.20	0.08
Hexi	0.04	0.66	0.22	0.10
Brahmaputra R.	0.05	0.66	0.25	0.09

Salween R.	0.08	0.66	0.27	0.10
Yangtze R.	0.04	0.67	0.27	0.11
Mekong R.	0.09	0.66	0.31	0.11
Yellow R.	0.05	0.67	0.34	0.12

3.6 Area of applicability and per-pixel uncertainty of the fused product

Figure 11 shows the fused DI and the four-model AOA intersection within the permafrost domain. Individual model AOAs cover 63.6% (XGBoost) to 77.1% (Random Forest) of the permafrost domain, and the four-model intersection—the fused AOA used as the final applicability mask—covers 59.9% (Table 5, Fig. S3). Outside the fused AOA, the DI exceeds the upper-whisker threshold of at least one model's training distribution, indicating extrapolation beyond the calibration domain; the excluded areas are concentrated in the higher-elevation portions of the central plateau and the driest parts of the Qaidam Basin, consistent with the geographic distribution of the training observations. In addition, the north–south-oriented red bands visible around ~95°E, 97°E and 99°E in the DI map (Fig. 11a) do not reflect genuine environmental dissimilarity but originate from a preprocessing artefact of the LST input layer: during the construction of the multi-year Landsat 8 ST_B10 maximum composite, cloud- and snow-contaminated pixels along these swaths were zero-filled rather than flagged as no-data, yielding anomalously low LST values. None of the 342 training samples coincides with the zero-fill pixels, so model calibration was unaffected; in the standardised predictor space, however, these anomalous LST values produce extreme DI, so the affected pixels are automatically flagged outside the AOA and are excluded when the product is masked to the fused AOA (Sect. 4.4; Fig. S3 of the Supplementary). Based on the DI–RMSE relationships established under the spatial cross-validation (Fig. S4), a per-pixel expected-RMSE layer was derived, increasing from approximately 0.05–0.06 $m^3\ m^{-3}$ at low DI to beyond 0.15 $m^3\ m^{-3}$ in the most dissimilar regions; together with the four-model prediction SD (mean = 0.024 $m^3\ m^{-3}$), these companion layers allow users to gauge the local reliability of the product. Users are advised to mask the product to the fused AOA when interpreting pixel values in environmentally novel regions.

Table 5:AOA coverage of each model and the four-model intersection, expressed relative to the QTP permafrost region.

Model	AOA coverage (% of permafrost region)	DI threshold
Random Forest	77.1	1.99
CatBoost	64.9	2.02
XGBoost	63.6	1.7
Extra Trees	75	2.01
Intersection	59.9	ND

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

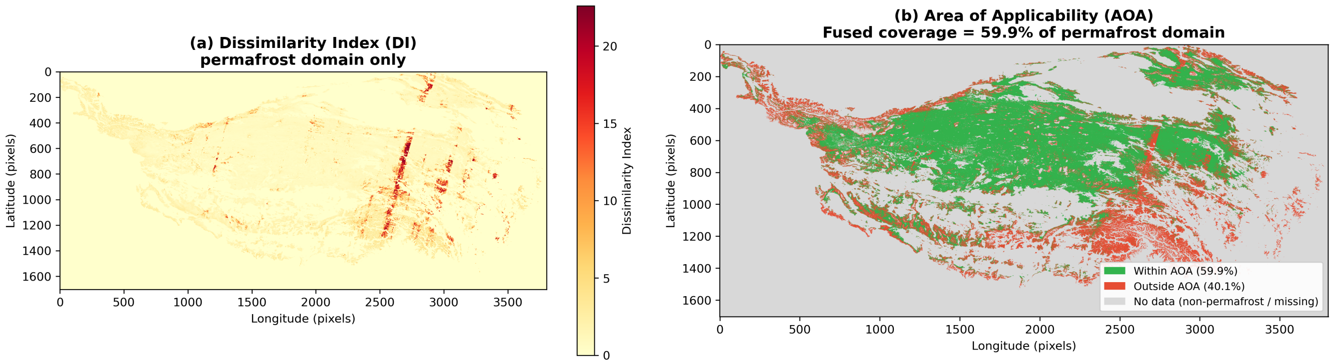

Figure 11: Fused dissimilarity index (DI) and area of applicability (AOA) within the QTP permafrost domain: (a) DI distribution; (b) fused AOA mask (green: within AOA, 59.9% of the permafrost domain; red: outside AOA; grey: no data / non-permafrost).

4 Discussion

4.1 Dominant Controls on Active Layer Moisture Distribution

The consistent identification of surface remote sensing factors (DeltaSigma, NDVI, NDWI, LST) as the primary drivers of ALM across all four models (combined contribution: 58–69%) reflects the strong coupling between surface hydrothermal states and subsurface moisture conditions in the active layer. The Sentinel-1 DeltaSigma index captures the seasonal contrast in microwave backscatter between wet (June) and frozen dry (January–February) conditions; its high explanatory power (mean SHAP contribution: 16–26% across models) indicates that the amplitude of seasonal surface wetness change is a robust proxy for ALM storage capacity. This is physically consistent with the role of soil moisture in controlling the dielectric properties of near-surface soil, which in turn governs C-band microwave backscatter intensity (Du et al., 2016; Xing et al., 2025). NDVI reflects vegetation abundance and productivity, both of which are tightly linked to soil moisture availability in the water-limited environments of the QTP (Li et al., 2024; Wang et al., 2016). The negative relationship between LST and ALM corroborates the well-established soil moisture–land surface temperature coupling mediated by evapotranspiration (Sandholt et al., 2002; Zhou et al., 2023).

Among soil physicochemical properties, OC_0–5 was the most important predictor in the three models that retained it (SHAP rank: 5th in CatBoost, 7th in Extra Trees, and 11th in XGBoost), consistent with the known role of organic matter in enhancing soil water retention through its effects on pore structure, aggregate stability, and hygroscopic properties. This finding aligns with observations from other permafrost regions where high soil organic carbon content is associated with elevated soil moisture (Schuur et al., 2015). The significant negative correlation between ALM and soil pH (Spearman r = −0.53) likely reflects the co-variation of pH with organic matter content and soil texture, as organic-rich and fine-textured soils in the humid eastern QTP tend to be more acidic. Topographic factors, while secondary in aggregate contribution, play a particularly significant modulating role in XGBoost predictions—where terrain-driven moisture redistribution processes

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

(slope gradient, relative slope position, flow accumulation) account for up to 14.5% of predicted ALM variability—consistent with the topographic control of lateral subsurface flow and water convergence in permafrost hillslopes (Walvoord and Kurylyk, 2016).

4.2 Multi-Model Fusion and Systematic Bias Correction

The systematic "low-value overestimation, high-value underestimation" bias pattern observed across all four models is a well-recognised artefact in machine learning-based soil property mapping. It arises primarily from sample distribution imbalance—training data are concentrated in the intermediate ALM range (0.2–0.4 $m^3\ m^{-3}$) with relatively few samples at the distribution tails—combined with the intrinsic tendency of ensemble tree models toward mean regression (Belitz and Stackelberg, 2021). Individual model architecture differences further contribute to spatial heterogeneity of bias: Boosting models (CatBoost, XGBoost) tend to produce more extreme dry predictions in arid endorheic basins, whereas Bagging models (RF, ET) produce more moderate estimates, reflecting their fundamentally different variance–bias trade-off strategies.

The proposed bias-aware fusion strategy—which separately applies minimum, maximum, and mean ensemble operations to the low (< 0.2 $m^3\ m^{-3}$), high (> 0.4 $m^3\ m^{-3}$), and intermediate ALM ranges, respectively—effectively leverages the complementary bias properties of different model architectures to correct systematic errors in a physically motivated manner. The improvement in the regression slope from 0.87–0.98 (individual models) to 0.99 (fused product) demonstrates the efficacy of this approach. This directional fusion strategy is particularly well suited for imbalanced training datasets where systematic prediction bias follows a monotonic pattern relative to the target variable, and may be applicable to other spatial soil property mapping tasks with similar characteristics.

The bias-aware fusion rule was motivated by a residual analysis showing that all four models systematically overestimate at low ALM and underestimate at high ALM under the spatial cross-validation. The R^2-weighted and simple-mean alternatives are nominally superior on the pooled hold-out metrics (R^2 = 0.387 and 0.383, respectively, versus 0.368 for the bias-aware rule; Table 3), but this margin is practically negligible (ΔRMSE ≤ 0.002 $m^3\ m^{-3}$, an order of magnitude below the retrieval uncertainty of the GPR reference data; Sect. 2.1) and comes at a mechanistic cost: mean-type averaging preserves the regression-to-the-mean tendency of the individual models, sustaining an underestimation of approximately 0.14 $m^3\ m^{-3}$ for pixels with ALM > 0.4 $m^3\ m^{-3}$ and compressing the spatial variance of the product. Because the objective of this study is a spatially continuous distribution dataset, in which faithful representation of the dry and wet tails matters as much as point-wise accuracy, the bias-aware rule was retained as the formal scheme: it corrects the diagnosed directional biases in a transparent and physically interpretable manner while preserving the full value range (0.02–0.70 $m^3\ m^{-3}$) of the ensemble predictions. The narrow spread of pooled R^2 across the five fusion rules (0.368–0.387, Table 3) confirms that the product is insensitive to the fusion rule. A sample-weighting remediation experiment (inverse-frequency weighting of the under-

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

represented low- and high-ALM ranges) provided only limited improvement, supporting the conclusion that the bias-aware fusion already mitigates the sample imbalance at the inference stage.

4.3 Dataset Evaluation and Comparison with Existing SSM Products

In our study, all four ensemble tree-based models yielded consistent cross-validated performance (R^2 = 0.62–0.63, RMSE = 0.08 $m^3\ m^{-3}$, MAE = 0.06 $m^3\ m^{-3}$). The close agreement across independently trained models demonstrates that prediction skill is robust rather than model-specific. The most directly comparable existing products are restricted to the Surface Soil Moisture (SSM) near-surface (0–5 cm) layer. In permafrost regions, ALM and near-surface SSM are typically highly correlated (areas with higher ALM generally coincide with higher SSM, and vice versa); comparing the spatial distribution of our ALM product with the existing SSM products can therefore corroborate the plausibility of the overall spatial pattern of the ALM retrieved in this study. Li et al. (2025) generated 100 m resolution SSM for the QTP thawing season; validation against 86 in situ monitoring stations yielded r = 0.72 and RMSE = 0.07–0.08 $m^3\ m^{-3}$. Xing et al. (2025) retrieved 1 km SSM from Sentinel-1 SAR, reporting median R = 0.57, ubRMSE = 0.064 $m^3\ m^{-3}$, RMSD = 0.107 $m^3\ m^{-3}$, and bias = 0.042 $m^3\ m^{-3}$ against independent observations. The present study achieves RMSE = 0.08 $m^3\ m^{-3}$ on par with these surface retrieval products, despite targeting the full active layer profile—a considerably more challenging modelling objective, given that direct microwave sensitivity to soil moisture decays rapidly below ~5 cm depth and that in situ calibration data for subsurface layers are far sparser than surface observations.

Figure S5a–c shows the spatial distribution of the fused ALM product, the Li SSM, and the Xing SSM over the permafrost domain. All three maps reproduce the well-documented southeast-to-northwest moisture gradient: high values in the Yellow River and Yangtze River headwaters (0.27–0.34 $m^3\ m^{-3}$), intermediate values in the central plateau (0.15–0.25 $m^3\ m^{-3}$), and low values in the Qaidam Basin and the northwestern Qiangtang Plateau (< 0.15 $m^3\ m^{-3}$). The pixel-level spatial correlation between ALM and Li SSM is r = 0.73 ($n \approx 1.6 \times 10^8$), confirming that the ALM product captures the same large-scale hydro-climatic gradients as the independent SSM retrieval. The Xing SSM covers only ~60 % of the permafrost domain in our study period, with its valid pixels concentrated in the southeastern and central plateau where Sentinel-1 backscatter constraints are most favorable; within these areas the spatial pattern is also broadly consistent with ALM.

Difference maps (Fig. S5d, e) and scatter plots (Fig. 12) quantify the systematic offsets between ALM and the SSM products.

The density scatter plot in Fig. 12b reveals two principal clusters separated by a relative minimum near ALM ≈ 0.18–0.20 $m^3\ m^{-3}$. This separation is a structural feature of the bias-aware fusion strategy rather than an artefact of model overfitting or input-grid defects: below the 0.2 $m^3\ m^{-3}$ threshold the fused value is progressively pulled toward the minimum of the four model predictions, and above 0.4 $m^3\ m^{-3}$ toward the maximum (Sect. 2.9.3), so comparatively few pixels obtain intermediate fused values in the transition band near 0.2 $m^3\ m^{-3}$. The ±0.05 $m^3\ m^{-3}$ symmetric linear transition windows imposed around each threshold smooth this transition and prevent abrupt histogram discontinuities (Figs. 8 and 9). The residual dip at ≈ 0.18–

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

0.20 m³ m⁻³ therefore reflects the rarity of pixels whose four-model mean falls within the low-ALM transition window, not a data or model defect. The spatial consistency between the ALM product and independent SSM retrievals (r = 0.73 for Li SSM) corroborates that the overall predicted spatial pattern is physically meaningful.

Because ALM integrates moisture throughout the entire active layer (typically tens of centimeters to several meters), it is systematically higher than the 0–5 cm Li SSM: mean ALM = 0.205 m³ m⁻³ versus mean Li SSM = 0.125 m³ m⁻³ over the permafrost domain, with a mean bias of +0.077 m³ m⁻³. The bias varies with moisture level: it is largest in the driest regions (0.13 m³ m⁻³ for SSM < 0.05 m³ m⁻³), decreases to ~0.05 m³ m⁻³ in the intermediate range (0.10–0.15 m³ m⁻³), and rises again to ~0.13 m³ m⁻³ in the wettest range (0.25–0.30 m³ m⁻³). This ALM-dependent structure reflects the nonlinear increase in moisture with depth—drier soils have a thinner saturated zone, while wetter soils sustain deeper percolation and thus larger depth-integrated values. Against the Xing SSM, the mean bias is –0.044 m³ m⁻³, but the spatial correlation is weak (r = 0.10) because the Xing product covers a more limited and spatially fragmented domain (59.6 % of the permafrost zone) and its 1 km resolution smooths local heterogeneity that is resolved in the 90 m ALM product.

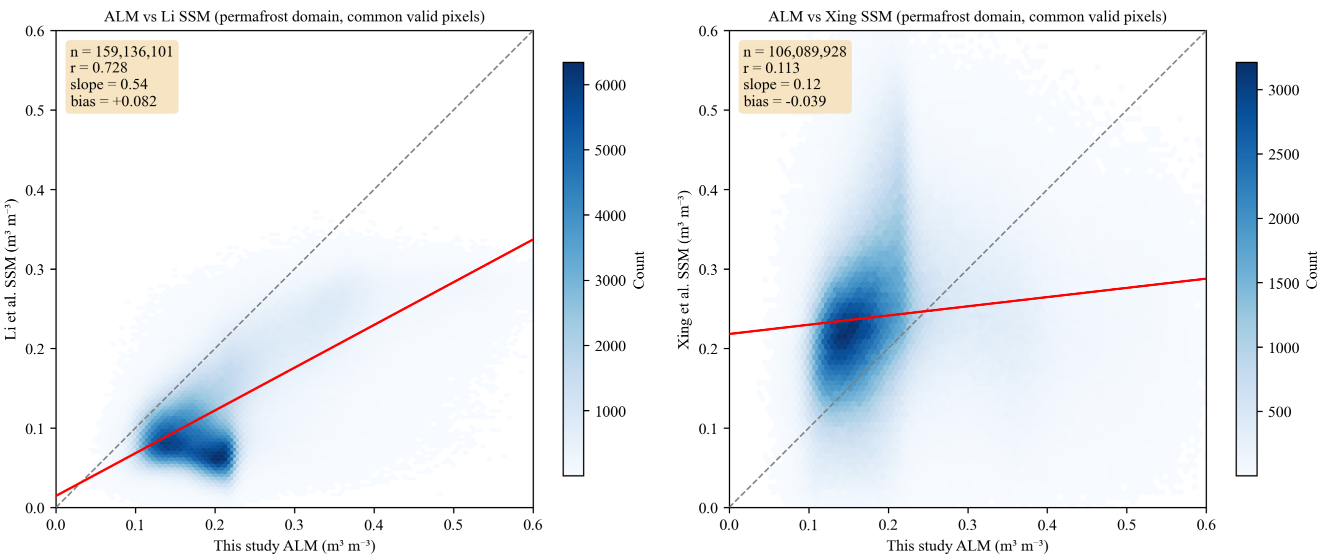

Figure 12: Density scatter plots of ALM versus (a) Li SSM and (b) Xing SSM within the common valid pixels of the permafrost domain. The grey dashed line is the 1:1 reference; the red line is the linear regression fit.

Three principal sources of uncertainty limit the achievable accuracy and should be considered when interpreting the dataset. First, the 342 in situ ALM samples are spatially concentrated along accessible terrain in the central plateau, with sparse coverage in the western plateau and permafrost transition zones, introducing spatial extrapolation uncertainty. Second, the training samples span 2009–2024, such that interannual variability in soil moisture and land surface conditions introduces noise into predictor–target relationships that temporally fixed remote sensing composites cannot fully resolve; the dataset should thus be interpreted as a climatological representation of peak-thaw season ALM rather than a snapshot for any

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

specific year. Third, the reliance on surface-sensitive remote sensing predictors (Sentinel-1 DeltaSigma, NDVI, LST) for inferring subsurface ALM means that model skill for deeper layers is mediated through the correlation structure between surface signals and stable terrain and soil physical attributes—a proxy relationship that may break down in areas with atypical soil profiles or unusual local hydrological connectivity. Notwithstanding these constraints, the convergence of R^2, RMSE, and MAE across the four independently trained models, together with the bias-aware multi-model fusion, provide confidence that the dataset captures the primary spatial patterns of ALM at the plateau scale.

4.4 Spatial transferability and area of applicability

The0.20–0.33drop in R^2 between random and spatial cross-validation (Table 2) exceeds the 0.10–0.20 range typically reported in the spatial-prediction literature (Ploton et al., 2020; Wadoux et al., 2021), indicating that the ALM–environment relationship is strongly modulated by basin-specific factors not captured by the current predictors. A Ploton-style diagnostic confirmed that the model is not simply exploiting spatial proximity: a Random Forest trained on sample coordinates alone achieved a spatial-CV R^2 of −0.19, far below the 0.30 obtained by the full-feature model, and a coordinate-inversion experiment (using the predictors to estimate sample locations) yielded a two-dimensional localisation error of approximately 250 km, indicating that the predictor combination does not uniquely identify individual samples. The random-CV optimism is therefore not primarily driven by spatial autocorrelation, but rather by the inability of the spatial folds to reproduce the strong east–west climate gradient within the training set—consistent with the prominent contribution of the PET predictor (Sect. 3.2). We therefore frame the spatial-CV results as a conservative lower bound of transferability, complemented by the AOA analysis that delineates where the random-CV performance can be expected to hold.

The four-model AOA intersection covers approximately 60% of the QTP permafrost region (Table 5, Fig. 11). Outside the fused AOA, at least one model extrapolates beyond its calibration domain, and the per-pixel expected RMSE derived from the DI–RMSE relationships increases accordingly. Users of the product are advised to mask the data to the AOA before interpreting pixel values in environmentally novel regions; the AOA mask and the per-pixel uncertainty layers are provided as companion layers in the dataset archive, clipped to the permafrost zone for consistency with the ALM product.

4.5 Limitations and Future Perspectives

Several limitations of this study deserve attention. First, the training sample size (n = 342) is moderate relative to the spatial extent of the QTP permafrost region (approximately 1.50×10^6 km²), and the samples are not uniformly distributed—density is higher along roads and accessible terrain, which may contribute to systematic biases at extreme ALM values and limit generalization to remote, under-sampled regions. Future campaigns should prioritize expanding sampling in the Qiangtang Plateau interior and high-mountain areas where observational data are sparsest.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Second, the temporal representativeness of the product requires consideration. Active layer moisture observations span 2009–2024 and represent the peak-thaw season (September–October), while remote sensing covariates are compiled as multi-year composites over 2014–2024, potentially introducing temporal mismatches between predictors and observations.

Third, the product represents column-integrated ALM, i.e., the depth-averaged moisture content of the entire active layer. Depth-resolved moisture profiles would provide richer information for hydrothermal modelling but would require substantially larger observational datasets at multiple depths—a key target for future field campaigns.

Fourth, as climate warming continues to drive permafrost degradation on the QTP, the moisture regime of the active layer is expected to undergo significant changes through altered precipitation patterns, increased evapotranspiration, and active layer thickening. Time-series modelling of active layer moisture change will be an important direction for future research.

5 Code and Data Availability

The 90 m resolution ALM dataset for the QTP permafrost region produced in this study is publicly available at the National Cryosphere Desert Data Center (NCDDC) with DOI: https://doi.org/10.12072/ncdc.permafrost.db7312.2026 (Du, 2026). The dataset is provided in GeoTIFF format and can be used with GIS software. The file contains a single band representing mean peak-thaw season ALM ($m^3\ m^{-3}$) for the QTP permafrost region. Ancillary datasets used in the analysis, including the permafrost boundary mask and basin delineations, are available from the original cited sources. All remote sensing input data are publicly accessible via the Google Earth Engine platform (https://earthengine.google.com), with dataset identifiers provided in Sects. 2.2–2.5. The 342 field-surveyed ALM observations are openly provided with DOI: https://www.doi.org/10.12072/ncdc.permafrost.db7703.2026 (Du, 2026).

The complete processing and modelling workflow—including the Spearman correlation analysis, the triple cross-validation SHAP-RFE feature selection for the four algorithms (RF, CatBoost, XGBoost, and Extra Trees), the grouped spatial cross-validation, the bias-aware multi-model fusion, the AOA/DI calculation, and all scripts reproducing the figures and tables of this paper—is publicly archived on Zenodo (https://doi.org/10.5281/zenodo.21999365). The Google Earth Engine script used to derive the Sentinel-1 backscatter difference (DeltaSigma) covariate is included in the same archive. Ground-penetrating radar data were processed with the commercial software Reflexw, and the resulting ALM observations are included in the published training dataset (https://doi.org/10.12072/ncdc.permafrost.db7703.2026).

6 Conclusions

This study presents the first 90 m resolution, spatially continuous dataset of ALM for the entire permafrost region of the Qinghai–Tibet Plateau (QTP), constructed by fusing 342 field-observed ALM samples with multi-source remote sensing, topographic, and soil physicochemical covariates using four ensemble tree-based machine learning models (CatBoost, Extra

Trees, Random Forest, and XGBoost) within a nested cross-validation and SHAP-RFE framework. The key conclusions and contributions are as follows:

(1) Surface remote sensing factors—particularly the Sentinel-1 backscatter difference index (DeltaSigma), NDVI, NDWI, and Landsat LST—are the primary drivers of ALM spatial variability. Soil organic carbon (OC_0–5) consistently emerges as the most important non-remote-sensing predictor, reflecting the critical role of organic matter in soil water retention capacity.

(2) All four ensemble tree-based models (Random Forest, Extra Trees, XGBoost, CatBoost) achieve comparable cross-validation performance (R^2 = 0.62–0.63, RMSE ≈ 0.08 $m^3 m^{-3}$), and exhibit consistent systematic biases: overestimation at low ALM (< 0.2 $m^3 m^{-3}$) and underestimation at high ALM (> 0.4 $m^3 m^{-3}$).

(3) The proposed fusion strategy—which applies minimum-, maximum-, and mean-ensemble operations to the low, high, and intermediate ALM ranges, respectively—effectively corrects systematic directional biases of individual models. The fused product achieves a regression slope of0.99, representing a substantial improvement in overall accuracy.

(4) The ALM for the QTP permafrost zone ranges from 0.02 to 0.70 $m^3 m^{-3}$, with a mean of 0.21 $m^3 m^{-3}$ and SD = 0.09 $m^3 m^{-3}$. A dominant southeast-high, northwest-low gradient is observed, with the driest conditions concentrated in the Qiangtang Plateau endorheic lake basins (mean ALM = 0.15–0.18 $m^3 m^{-3}$) and the highest moisture in the Yellow River and Mekong River headwaters (mean ALM = 0.31–0.34 $m^3 m^{-3}$). Higher-ALM regions consistently exhibit greater spatial variability. An area-of-applicability analysis further indicates that the cross-validated performance can be expected to hold over approximately 60% of the permafrost domain, and per-pixel uncertainty layers are provided to support reliability-aware applications.

This dataset fills a critical data gap in full-depth active layer soil moisture mapping at regional scales on the QTP, and provides a solid data foundation for advancing research on permafrost hydrothermal dynamics, land–atmosphere interactions, hydrological modeling, ecological assessment, and climate change impact evaluation in the Third Pole region.

Author contributions

E. Du conceptualized the study, conducted the formal analysis, led the field investigations, developed the methodology, wrote the software, performed the visualization, and prepared the original draft. Y. Min, Y. Ran, and K. Feng assisted with data curation. Y. Ran also contributed to conceptualization. Z. Xing contributed to formal analysis. Y. Xiao and Z. Li assisted with methodology development and software implementation. D. Zou, G. Liu, X. Zhu, and J. Chen participated in field investigations. L. Zhao, L. Dai, J. Yao, L. Wang, and G. Hu contributed to conceptualization. T. Wu, L. Dai, L. Zhao, G. Hu, and J. Chen administered the project. L. Zhao, T. Wu, G. Hu, and Y. Min acquired funding. J. Yao, T. Wu, and L.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Zhao provided supervision. L. Zhao, T. Wu, L. Wang, J. Yao, G. Hu, and Y. Ran reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The authors gratefully acknowledge Prof. Xiahua Hao (Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences) and Dr. Qin Zhao (School of Vehicle and Transportation Engineering, Taiyuan University of Science and Technology) for their constructive advices during the modelling process. The authors also thank Engineer Chenchen Tian and Engineer Wenlong Ma (Institute of Earth Environment, Chinese Academy of Sciences) for their technical support in remote sensing image processing on the Google Earth Engine (GEE) platform.

Financial support

This study was jointly supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant no: XDB0950000) and the National Natural Science Foundation of China projects (Grant no: 42525105), Additional support was provided by the Open Foundation of the National Cryosphere Desert Data Center (Grant no. 2025NCDC010).

Review statement

The review statement will be added by Copernicus Publications listing the handling editor as well as all contributing referees according to their status anonymous or identified.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

References

Bauer-Marschallinger, B., Cao, S., Navacchi, C., Freeman, V., Reuß, F., Geudtner, D., Rommen, B., Vega, F. C., Snoeij, P., Attema, E., Reimer, C., and Wagner, W.: The normalised Sentinel-1 Global Backscatter Model, mapping Earth's land surface with C-band microwaves, Sci. Data, 8, 277, https://doi.org/10.1038/s41597-021-01059-7, 2021.

Belitz, K. and Stackelberg, P. E.: Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models, Environ. Model. Softw., 139, 105006, https://doi.org/10.1016/j.envsoft.2021.105006, 2021.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cawley, G. C. and Talbot, N. L. C.: On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, J. Mach. Learn. Res., 11, 2079–2107, 2010.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Cheng, G., Zhao, L., Li, R., Wu, X., Sheng, Y., Hu, G., Zou, D., Jin, H., Li, X., and Wu, Q.: Characteristic, changes and impacts of permafrost on Qinghai-Tibet Plateau (in Chinese), Chin. Sci. Bull., 64, 2783–2795, https://doi.org/10.1360/TB-2019-0191, 2019.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., De Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., Van Der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sens. Environ., 203, 185–215, https://doi.org/10.1016/j.rse.2017.07.001, 2017.

Dorogush, A. V., Ershov, V., and Gulin, A.: CatBoost: gradient boosting with categorical features support, [preprint], https://doi.org/10.48550/ARXIV.1810.11363, 2018.

Du, E.: 90 m Active Layer Moisture Dataset for the Qinghai–Tibet Plateau Permafrost Region (2009-2024), NCDC [dataset], https://doi.org/10.12072/ncdc.permafrost.db7312.2026, 2026a.

Du, E.: Sample data set of moisture survey in the active layer of the permafrost region of the Qinghai-Tibet Plateau (2009-2024), NCDC [dataset], https://doi.org/10.12072/ncdc.permafrost.db7703.2026, 2026b.

Du, E., Zhao, L., Wu, T., Li, R., Yue, G., Wu, X., Li, W., Jiao, Y., Hu, G., Qiao, Y., Wang, Z., Zou, D., and Liu, G.: The relationship between the ground surface layer permittivity and active-layer thawing depth in a Qinghai–Tibetan Plateau permafrost area, Cold Reg. Sci. Technol., 126, 55–60, https://doi.org/10.1016/j.coldregions.2016.03.006, 2016.

Du, E., Zhao, L., Zou, D., Li, R., Wang, Z., Wu, X., Hu, G., Zhao, Y., Liu, G., and Sun, Z.: Soil Moisture Calibration Equations for Active Layer GPR Detection—a Case Study Specially for the Qinghai–Tibet Plateau Permafrost Regions, Remote Sensing, 12, 605, https://doi.org/10.3390/rs12040605, 2020.

Du, R., Peng, X., Frauenfeld, O. W., Jin, H., Wang, K., Zhao, Y., Luo, D., and Mu, C.: Quantitative Impact of Organic Matter and Soil Moisture on Permafrost, J. Geophys. Res. Atmos., 128, e2022JD037686, https://doi.org/10.1029/2022JD037686, 2023.

Gao, B.: NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sens. Environ., 58, 257–266, https://doi.org/10.1016/S0034-4257(96)00067-3, 1996.

GCOS: Essential Climate Variables (ECVs) – Permafrost, World Meteorological Organization, Geneva, 2022.

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, Mach. Learn., 63, 3–42, https://doi.org/10.1007/s10994-006-6226-1, 2006.

Han, Q., Zeng, Y., Zhang, L., Wang, C., Prikaziuk, E., Niu, Z., and Su, B.: Global long term daily 1 km surface soil moisture dataset with physics informed machine learning, Sci. Data, 10, 101, https://doi.org/10.1038/s41597-023-01991-w, 2023.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X.,Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hinkel, K. M. and Nelson, F. E.: Spatial and temporal patterns of active layer thickness at Circumpolar Active Layer Monitoring (CALM) sites in northern Alaska, 1995–2000, J. Geophys. Res., 108, 2001JD000927, https://doi.org/10.1029/2001JD000927, 2003.

Huisman, J. A., Hubbard, S. S., Redman, J. D., and Annan, A. P.: Measuring Soil Water Content with Ground Penetrating Radar: A Review, Vadose Zone J., 2, 476–491, https://doi.org/10.2136/vzj2003.4760, 2003.

IPCC: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1st ed., Cambridge University Press, https://doi.org/10.1017/9781009157896, 2023.

Ji, F., Fan, L., Kuang, X., Li, X., Cao, B., Cheng, G., Yao, Y., and Zheng, C.: How does soil water content influence permafrost evolution on the Qinghai-Tibet Plateau under climate warming?, Environ. Res. Lett., 17, 064012, https://doi.org/10.1088/1748-9326/ac6c9a, 2022.

Jin, H.-J., Wu, Q.-B., and Romanovsky, V. E.: Degrading permafrost and its impacts, Adv. Clim. Change Res., 12, 1–5, https://doi.org/10.1016/j.accre.2021.01.007, 2021.

Li, W., Yan, D., Weng, B., Lai, Y., Zhu, L., Qin, T., Dong, Z., and Bi, W.: Nonlinear effects of surface soil moisture changes on vegetation greenness over the Tibetan plateau, Remote Sens. Environ., 302, 113971, https://doi.org/10.1016/j.rse.2023.113971, 2024.

Li, Z., Zhao, L., Wang, L., Liu, G., Du, E., Zou, D., Hu, G., Xing, Z., Wang, C., Liu, S., Xiao, M., Yin, L., and Wang, Y.: 100-m-resolution surface soil moisture data during the thawing season on the Qinghai–Tibet Plateau, Sci. Data, 12, 497, https://doi.org/10.1038/s41597-025-04831-1, 2025.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 30, 2017.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 4540, https://doi.org/10.1038/s41467-020-18321-y, 2020.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Sandholt, I., Rasmussen, K., and Andersen, J.: A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status, Remote Sens. Environ., 79, 213–224, https://doi.org/10.1016/S0034-4257(01)00274-7, 2002.

Schuur, E. A. G., McGuire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., Hugelius, G., Koven, C. D., Kuhry, P., Lawrence, D. M., Natali, S. M., Olefeldt, D., Romanovsky, V. E., Schaefer, K., Turetsky, M. R., Treat, C. C., and Vonk, J. E.: Climate change and the permafrost carbon feedback, Nature, 520, 171–179, https://doi.org/10.1038/nature14338, 2015.

Shi, G., Sun, W., Shangguan, W., Wei, Z., Yuan, H., Li, L., Sun, X., Zhang, Y., Liang, H., Li, D., Huang, F., Li, Q., and Dai, Y.: A China dataset of soil properties for land surface modelling (version 2, CSDLv2), Earth Syst. Sci. Data, 17, 517–543, https://doi.org/10.5194/essd-17-517-2025, 2025.

Varma, S. and Simon, R.: Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics, 7, 91, https://doi.org/10.1186/1471-2105-7-91, 2006.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., De Bruin, S., and Brus, D. J.: Spatial cross-validation is not the right way to evaluate map accuracy, Ecol. Modell., 457, 109692, https://doi.org/10.1016/j.ecolmodel.2021.109692, 2021.

Walvoord, M. A. and Kurylyk, B. L.: Hydrologic Impacts of Thawing Permafrost—A Review, Vadose Zone J., 15, 1–20, https://doi.org/10.2136/vzj2016.01.0010, 2016.

Wang, X., Yi, S., Wu, Q., Yang, K., and Ding, Y.: The role of permafrost and soil water in distribution of alpine grassland and its NDVI dynamics on the Qinghai-Tibetan Plateau, Glob. Planet. Change, 147, 40–53, https://doi.org/10.1016/j.gloplacha.2016.10.014, 2016.

Wlostowski, A. N., Gooseff, M. N., and Adams, B. J.: Soil Moisture Controls the Thermal Habitat of Active Layer Soils in the McMurdo Dry Valleys, Antarctica, J. Geophys. Res. Biogeo., 123, 46–59, https://doi.org/10.1002/2017JG004018, 2018.

Xing, Z., Zhao, L., Fan, L., De Lannoy, G., Bai, X., Liu, X., Peng, J., Frappart, F., Yang, K., Li, X., Zhou, Z., Li, X., Zeng, J., Zou, D., Du, E., Wang, C., Wang, L., Li, Z., and Wigneron, J.-P.: Retrieval of 1 km surface soil moisture from Sentinel-1 over bare soil and grassland on the Qinghai-Tibetan Plateau, Remote Sens. Environ., 318, 114563, https://doi.org/10.1016/j.rse.2024.114563, 2025.

https://doi.org/10.5194/essd-2026-330
Preprint. Discussion started: 24 August 2026

Yao, T., Bolch, T., Chen, D., Gao, J., Immerzeel, W., Piao, S., Su, F., Thompson, L., Wada, Y., Wang, L., Wang, T., Wu, G., Xu, B., Yang, W., Zhang, G., and Zhao, P.: The imbalance of the Asian water tower, Nat. Rev. Earth Environ., 3, 618–632, https://doi.org/10.1038/s43017-022-00299-4, 2022.

Zhang, Y., Li, B., Liu, L., and Zheng, D.: Redetermine the region and boundaries of Tibetan Plateau (in Chinese), Geogr. Res., 40, 1543–1553, https://doi.org/10.11821/dlyj020210138, 2021.

Zhou, J., Yang, K., Crow, W. T., Dong, J., Zhao, L., Feng, H., Zou, M., Lu, H., Tang, R., and Jiang, Y.: Potential of remote sensing surface temperature- and evapotranspiration-based land-atmosphere coupling metrics for land surface model calibration, Remote Sens. Environ., 291, 113557, https://doi.org/10.1016/j.rse.2023.113557, 2023.

Zou, D., Zhao, L., Hu, G., Du, E., Liu, G., Wang, C., and Li, W.: Permafrost temperature baseline at 15 m depth on the Qinghai–Tibetan Plateau (2010–2019), Earth Syst. Sci. Data, 17, 1731–1742, https://doi.org/10.5194/essd-17-1731-2025, 2025.