Supplementary

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)

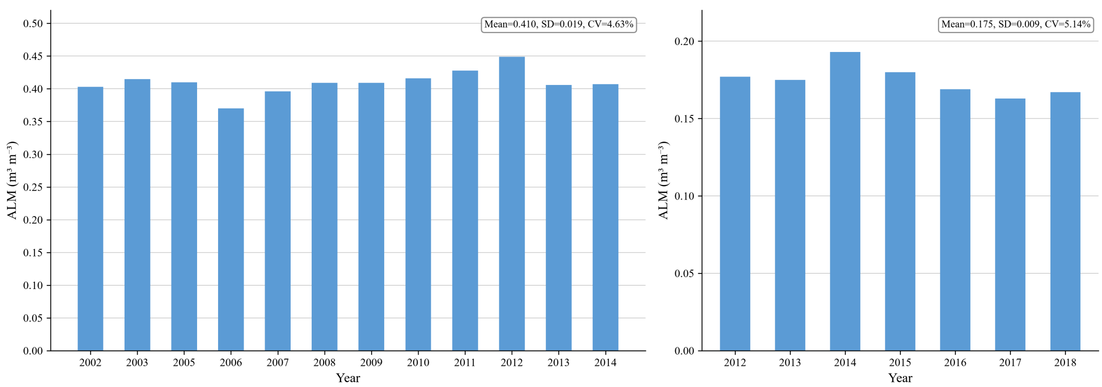

Figure S1: Interannual variability of active layer moisture at two monitoring sites along the Qinghai–Tibet Highway: swamp meadow at Liangdaohe (a) and alpine desert at Wudaoliang (b).

Figure S1a shows the interannual variability of active layer moisture recorded at the Liangdaohe monitoring site, located near the southern boundary of permafrost distribution along the Qinghai–Tibet Highway, during 2002–2013. The land surface at this site is characterized by typical swamp meadow vegetation, with an active layer thickness (ALT) of approximately 1.1 m (range: 1.1–1.2 m). Over the observation period, the mean active layer moisture (ALM) was 0.41 $m^3 m^{-3}$, with a standard deviation (SD) of 0.019 $m^3 m^{-3}$ and a coefficient of variation (CV, defined as the ratio of SD to mean) of 4.6%, indicating weak interannual variability.

Figure S1b shows the interannual variability of active layer moisture at the Wudaoliang monitoring site along the Qinghai–Tibet Highway during 2012–2018. The land surface at this site is characterized by typical alpine desert, with an ALT of approximately 2.4 m (range: 2.3–2.5 m). The mean ALM was 0.18 $m^3 m^{-3}$ (SD = 0.009 $m^3 m^{-3}$; CV = 5.1%), also indicating weak interannual variability.

For comparison, the 342 field-surveyed sample points used in this study yield a mean ALM of 0.24 $m^3 m^{-3}$, an SD of 0.13 $m^3 m^{-3}$, and a CV of 54.2%, reflecting pronounced spatial heterogeneity across the QTP permafrost region. Although interannual fluctuations in active layer moisture do exist at individual monitoring sites, these fluctuations are negligible relative to the spatial variability across the study area. This demonstrates that the use of field samples collected across different years (2009–2024) will not introduce substantial bias into the spatial prediction of active layer moisture in this study.

Table S1: Topographic factors used in this study: calculation formulae and physical interpretations.

Factor (unit)	**Formula**	**Definition and hydrological relevance**
Slope/degrees	$Slope=\frac{180}{\pi}\times\arctan\left(\sqrt{z_x^2+z_y^2}\right)$	The maximum rate of elevation change per unit horizontal distance at a given location, characterizing terrain steepness. Slope regulates moisture input, lateral redistribution, and retention processes, and is thus one of the most fundamental controls on soil moisture and near-surface hydrological patterns.
Aspect /degrees	$Aspect=\arctan 2\left(-z_x,-z_y\right)$	The azimuth of the steepest downslope direction of a surface facet, measured clockwise from north (0°) in a projected coordinate system. Aspect determines solar radiation receipt, snowmelt timing, evapotranspiration rates, and the directional redistribution of soil moisture.
Profile Curvature/ m^{-1}	$C_{profile}=\frac{-z_{xx}z_x^2+2z_{xy}z_x z_y+z_{yy}z_y^2}{\left(z_x^2+z_y^2\right)\left(1+z_x^2+z_y^2\right)^{3/2}}$	The curvature of the terrain surface along the direction of maximum slope gradient, characterising the rate of change of flow velocity. Negative values indicate concave (decelerating flow) surfaces; positive values indicate convex (accelerating flow) surfaces; zero indicates a planar surface.
Plan Curvature/m^{-1}	$C_{plan}=\frac{z_{xx}z_{yy}^2-2z_{xy}z_x z_y+z_{yy}z_x^2}{\left(z_x^2+z_y^2\right)^{3/2}}$	The curvature of contour lines, characterising the convergence or divergence of surface flow. Negative values indicate flow convergence; positive values indicate flow divergence; zero indicates rectilinear contours.
Topographic Contributing Area (TCA)/pixel	$TCA_i=1+\sum_{j\in U(i)} TCA_j$	The number of upslope grid cells that drain into a given cell, serving as a primary index of upslope flow accumulation capacity.
Topographic Wetness Index (TWI)/dimensionless	$TWI=\ln\left(\frac{a}{\tan\beta}\right)$	An integrated index reflecting the topographic control on soil moisture distribution. a is the specific catchment area ($m^2\ m^{-1}$) and β is the local slope angle (radians). Higher TWI values correspond to locations with greater potential for soil moisture accumulation.

Relative Slope Position (RSP)/dimensionless	$RSP = \frac{z_i - z_{min}}{z_{max} - z_{min}}$	A normalised index of a pixel's relative elevation position within the regional terrain. z_i, z_{min}, and z_{max} are the pixel elevation and the minimum and maximum elevations of the study area, respectively. RSP captures gravitational potential and long-term drainage conditions, and is a key factor explaining the background spatial pattern of moisture distribution.
LS Factor/ dimensionless	$LS = \left(\frac{a}{22.13}\right)^{0.5} \left(\frac{sin\beta}{0.0896}\right)^{1.3}$	A combined index of slope length (flow accumulation potential) and slope gradient (gravitational driving force), characterising the joint influence of topography on surface runoff and erosion processes. It serves as an important topographic constraint for explaining spatial heterogeneity in soil moisture.
Convergence Index (CI)/m^{-1}	$CI = -\left(z_{xx} + z_{yy}\right)$	A terrain index describing the tendency of surface flow to converge or diverge at the local scale. Positive CI values indicate divergent topography (e.g., ridgelines, hilltops); negative values indicate convergent topography (e.g., valleys, depressions, slope bases); values near zero indicate planar surfaces.
Valley Depth (VD)/m	$VD = z_{max}^{(W_r)} - z_i$	The vertical elevation difference between a given pixel and its surrounding local highland (e.g., ridgeline or hilltop), quantifying the degree of local topographic depression.

Note: In the equations, $z_x = \frac{\partial z}{\partial x}$ and $z_y = \frac{\partial z}{\partial y}$ represent the first-order derivatives of elevation in the east–west and north–south directions, respectively. $z_{xx} = \frac{\partial^2 z}{\partial x^2}$ and $z_{xy} = \frac{\partial^2 z}{\partial x \partial y}$ denote the second-order derivatives of elevation in the east–west and north–south directions, respectively, while $z_{xy} = \frac{\partial^2 z}{\partial x \partial y}$ represents the mixed second-order derivative in the orthogonal directions. $z_{max}^{(W_r)}$ denotes the maximum elevation within a local window centered at pixel i with radius r. In this study, the window size is defined as 11 × 11 pixels, corresponding to an approximate radius of 1 km.

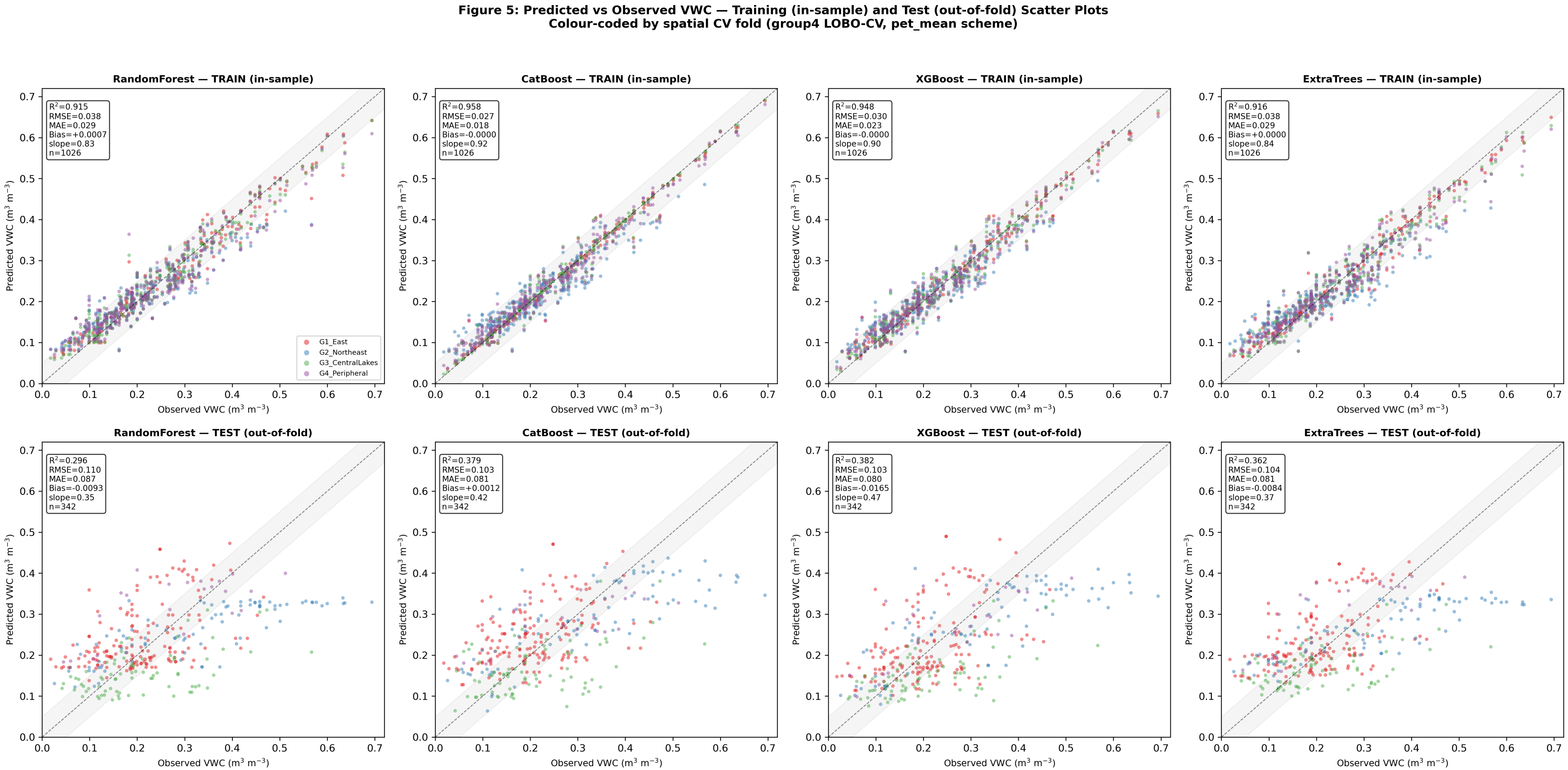

Figure S2: Observed versus predicted ALM scatter plots for (a) CatBoost, (b) Extra Trees, (c) Random Forest, and (d) XGBoost, pooling all training data (upper row) and all test data (lower row) across the five outer folds of spatial cross-validation. Points are colour-coded by spatial cross-validation fold. The dashed line indicates the 1:1 reference.

Figure S2 presents per-model observed-versus-predicted scatter plots corresponding to the pooled panels shown in Fig. 5 of the main manuscript. The upper row shows in-sample (training-set) predictions and the lower row shows out-of-fold (test-set) predictions across the five outer folds of group-based spatial cross-validation. Each point is colour-coded by spatial CV fold, and the dashed grey line marks the 1:1 reference. In-sample metrics are reported in the top-left corner of each upper panel: CatBoost (R^2 = 0.956, RMSE = 0.027 $m^3\ m^{-3}$, MAE = 0.021 $m^3\ m^{-3}$, bias = +0.001 $m^3\ m^{-3}$, slope = 0.93), XGBoost (R^2 = 0.949, RMSE = 0.030 $m^3\ m^{-3}$, MAE = 0.023 $m^3\ m^{-3}$, bias = –0.001 $m^3\ m^{-3}$, slope = 0.89), Extra Trees (R^2 = 0.918, RMSE = 0.038 $m^3\ m^{-3}$, MAE = 0.028 $m^3\ m^{-3}$, bias = +0.001 $m^3\ m^{-3}$, slope = 0.82), and Random Forest (R^2 = 0.915, RMSE = 0.038 $m^3\ m^{-3}$, MAE = 0.029 $m^3\ m^{-3}$, bias = +0.001 $m^3\ m^{-3}$, slope = 0.83). These high in-sample R^2 values (0.915–0.956) confirm that all four models memorise the training domain well.

Test-set metrics (lower row) reveal the models' spatial transferability limitations: CatBoost (R^2 = 0.379, RMSE = 0.081 $m^3\ m^{-3}$, slope = 0.62), XGBoost (R^2 = 0.382, RMSE = 0.080 $m^3\ m^{-3}$, slope = 0.47), Extra Trees (R^2 = 0.362, RMSE = 0.104 $m^3\ m^{-3}$, slope = 0.37), and Random Forest (R^2 = 0.296, RMSE = 0.087 $m^3\ m^{-3}$, slope = 0.35). The pronounced scatter (pooled test R^2 = 0.30–0.38) reflects the difficulty of extrapolating from calibrated basins to hydrologically and climatically distinct ungauged basins. A consistent systematic bias is visible in every lower panel: points lie above the 1:1 line at low ALM (< 0.2 $m^3\ m^{-3}$, overestimation) and below it at high ALM (> 0.4 $m^3\ m^{-3}$, underestimation). This bias pattern—rooted in the scarcity of extreme-value training samples—motivated the bias-aware fusion strategy adopted in the main manuscript (low-ALM pixels take the four-model minimum, high-ALM pixels take the maximum, and mid-range pixels take the mean).

Table S2: Final feature subsets retained by each model.

Model	**Number of features**	**Feature subset**
Random Forest	9	DeltaSigma, LST, NDVI, NDWI, AK_60–100, CEC_5–15, pH_5–15, porosity_15–30, pet_mean
CatBoost	8	DeltaSigma, LST, NDVI, NDWI, AP_5–15, OC_0–5, LS_factor, pet_mean
XGBoost	11	DeltaSigma, LST, NDVI, AK_60–100, AP_5–15, OC_0–5, porosity_15–30, Aspect_cos, LS_factor, RSP, pet_mean
Extra Trees	8	DeltaSigma, LST, NDVI, NDWI, CEC_5–15, OC_0–5, pH_5–15, pet_mean

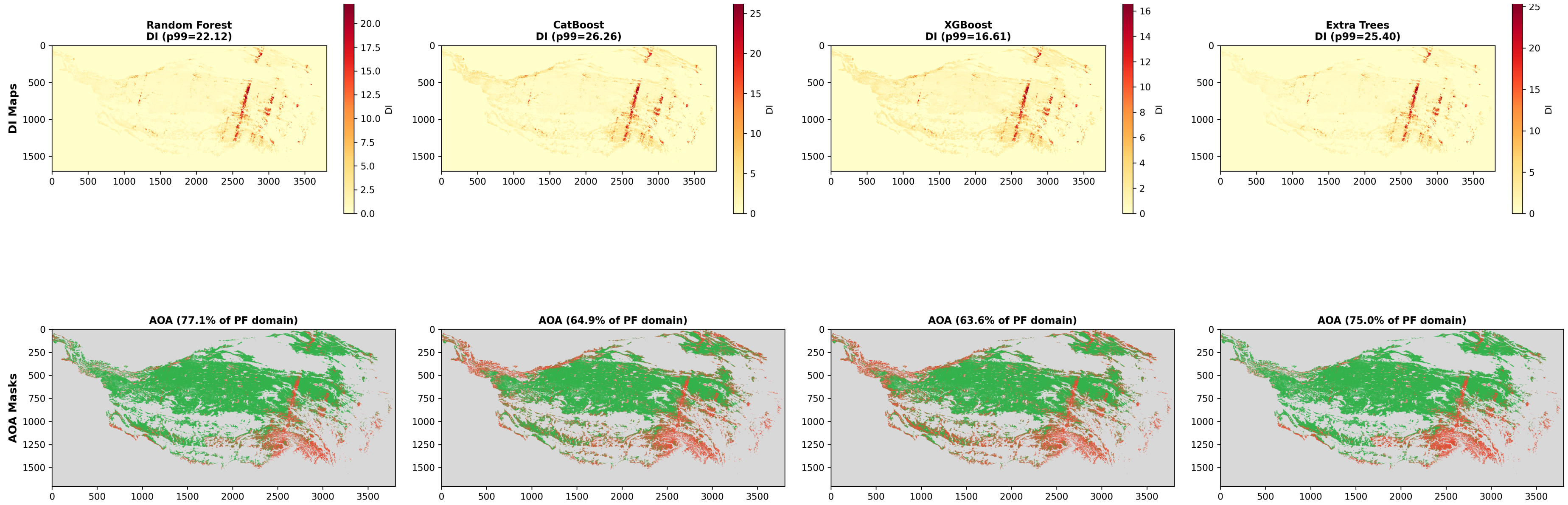

Figure S3: Per-model DI and AOA comparison within the permafrost domain for (a) CatBoost, (b) Extra Trees, (c) Random Forest, and (d) XGBoost (upper row: DI maps; lower row: AOA masks, with coverage percentages relative to the permafrost domain).

Figure S3 illustrates the per-model Area of Applicability (AOA) analysis within the QTP permafrost domain (Meyer & Pebesma, 2021). The upper row shows Dissimilarity Index (DI) maps for each model. DI quantifies the weighted Euclidean distance between each prediction pixel and its nearest training sample in the standardised feature space; higher DI (warm/red colours) indicates greater extrapolation distance and thus larger prediction uncertainty. Notable red striping along ~95°E, 97°E and 99°E primarily reflects a Landsat ST_B10 preprocessing artefact (zero-fill pixels in the LST input raster) rather than genuine environmental dissimilarity, as discussed in the main text (Sect. 3.6).

The lower row shows binary AOA masks: green denotes pixels inside the AOA (DI ≤ upper-whisker threshold, interpolation-conditional domain where cross-validated performance is expected to hold), and red denotes pixels outside the AOA (extrapolation domain, predictions should be used with caution). AOA coverage percentages within the permafrost domain differ across models—Random Forest 77.1%, Extra Trees 75.0%, CatBoost 64.9%, XGBoost 63.6%—because each model retains a distinct feature subset and therefore occupies a slightly different predictor space. Red (AOA-outside) areas are concentrated along the eastern permafrost margin (sparse training samples) and along the LST-defect strips, confirming that the AOA mask effectively flags locations where the models extrapolate beyond their empirical calibration envelope.

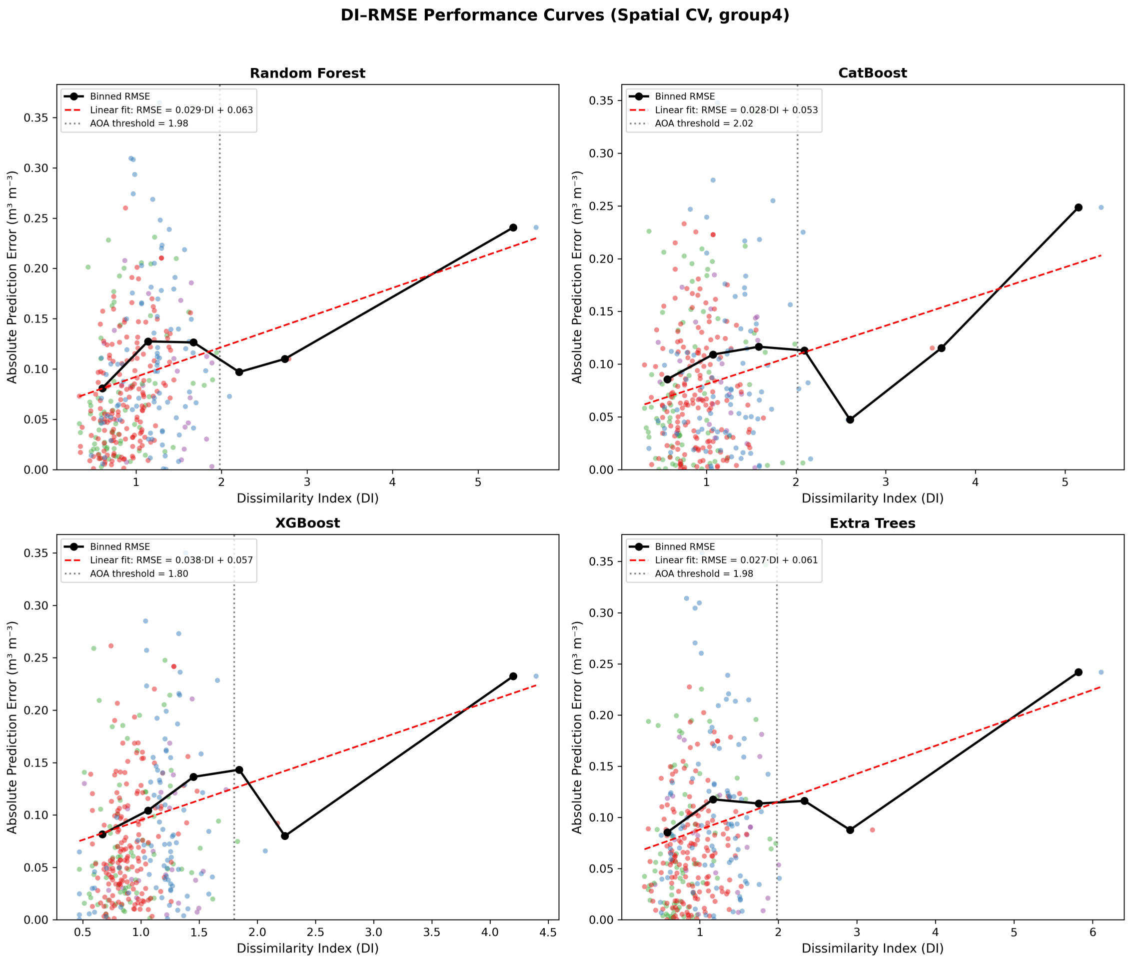

Figure S4: DI–RMSE performance curves under the spatial cross-validation for the four models: sample-level DI versus absolute prediction error, binned RMSE, and linear fits; grey dotted lines indicate the AOA DI thresholds.

Figure S4 validates the monotonic DI–RMSE relationship that underpins the AOA method. Each panel corresponds to one model. Coloured scatter points show, for every spatial-CV test sample, the pair (DI, absolute prediction error). The black solid line gives binned RMSE (mean absolute error within successive DI bins), and the red dashed line shows the linear fit reported in the legend (e.g., Random Forest: RMSE = 0.029·DI + 0.063; CatBoost: RMSE = 0.028·DI + 0.053; XGBoost: RMSE = 0.038·DI + 0.057; Extra Trees: RMSE = 0.027·DI + 0.061). The grey vertical dotted line marks the AOA threshold (upper whisker = Q3 + 1.5 × IQR of the training-sample DI distribution): Random Forest 1.98, CatBoost 2.02, XGBoost 1.80, Extra Trees 1.98.

All four models exhibit a clear positive DI–RMSE trend: as DI increases, binned RMSE rises systematically, confirming that pixels farther from the training domain in feature space incur larger prediction errors. Beyond the AOA threshold, the binned RMSE curves level off or fluctuate because of sparse high-DI samples, but the overall upward tendency remains. This empirical validation justifies using the AOA mask as a spatial proxy for prediction reliability: the green (AOA-inside) domain can be interpreted as the region where the cross-validated accuracy metrics reported in Table 2 are expected to hold, whereas the red (AOA-outside) domain should be treated as unquantified extrapolation.

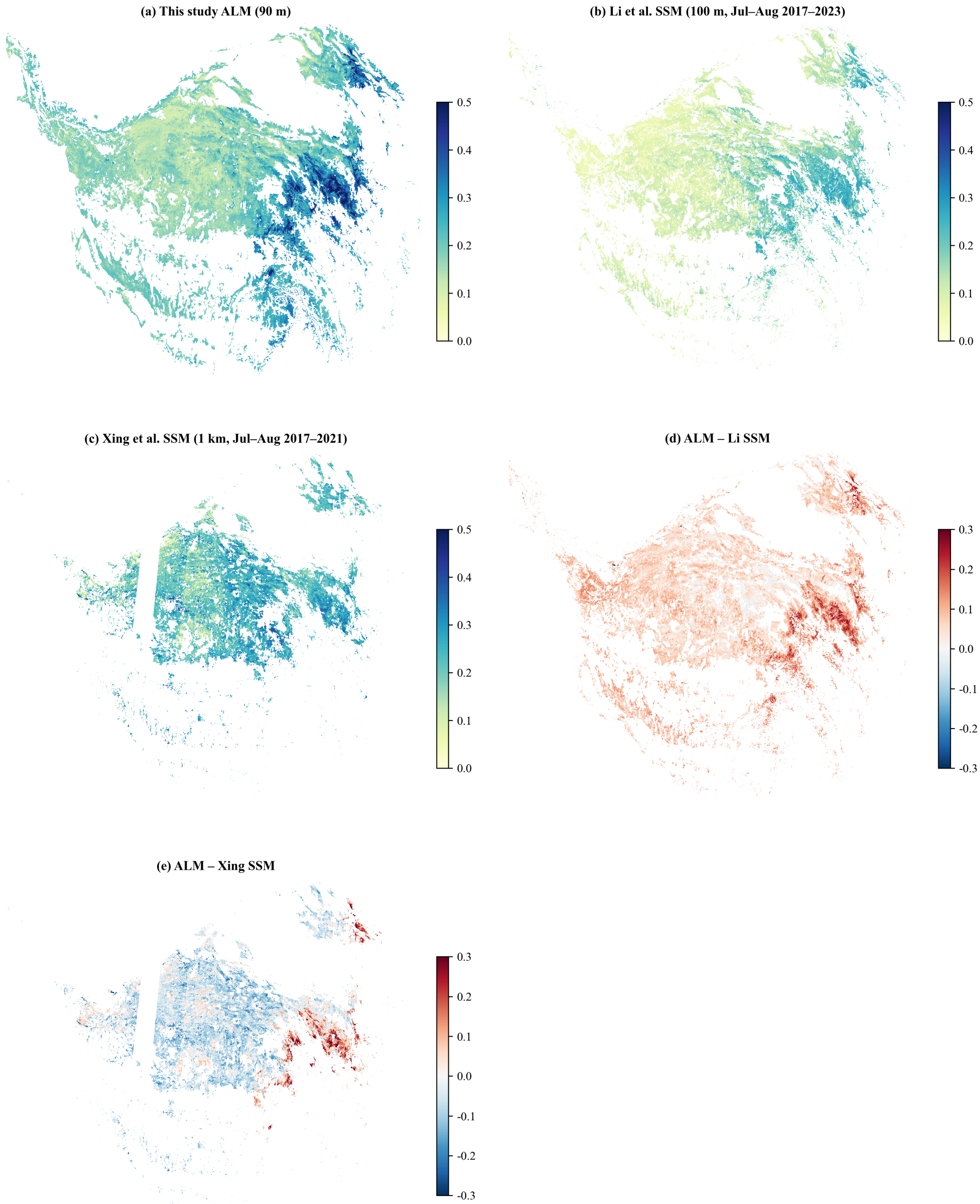

Figure S5: Spatial comparison of (a) this study ALM (90 m), (b) Li et al. SSM (100 m, July–August 2017–2023), (c) Xing et al. SSM (1 km, July–August 2017–2021), and difference maps (d) ALM – Li SSM and (e) ALM – Xing SSM over the QTP permafrost domain.

Figure S5 provides a spatial comparison between the final fused ALM product of this study (90 m, multi-year mean for June–September) and two independent surface soil moisture (SSM) reference datasets. Panel (b) shows Li et al. SSM (100 m, July–August mean, 2017–2023) and panel (c) shows Xing et al. SSM (1 km, July–August mean, 2017–2021). All three products display the same broad southeast-to-northwest moisture gradient across the QTP permafrost region, with dry conditions in the Qiangtang Plateau interior (Serling Co, Yibug Caka) and wetter conditions in the eastern river-source basins (Yellow River, Yangtze River, Mekong River).

Panel (d) maps ALM minus Li SSM differences. Red shading (positive differences) dominates the Qaidam Basin and the Qiangtang Plateau lake basins, where ALM is systematically higher than Li SSM, whereas blue shading (negative differences) appears in the eastern headwater regions. The overall Pearson correlation between ALM and Li SSM is $r = 0.73$ ($n \approx 1.6 \times 10^8$ common pixels), with a regression slope of 0.62 and a mean bias of $+0.089$ m^3 m^{-3} (ALM higher). The bias likely reflects physical differences in sensing depth (full active layer vs surface 0–5 cm) and seasonal window (June–September vs July–August), as well as the fact that SSM captures transient surface wetness whereas ALM integrates the deeper, more stable active-layer moisture reservoir.

Panel (e) maps ALM minus Xing SSM differences. The Xing product is coarser (1 km) and spatially sparser (many NoData pixels in the western plateau), so the overlap with ALM is patchier. The correlation is weaker ($r \approx 0.10$), largely because 1 km aggregation smooths out the fine-scale heterogeneity resolved by the 90 m ALM product, and because the Xing dataset's narrower temporal window (2017–2021) captures a different climatic envelope. Despite these discrepancies, the spatial pattern of wetter east / drier west is qualitatively consistent across all three datasets, supporting the overall plausibility of the ALM product.