ChinaAI-FSC: A Comprehensive AI-Ready MODIS Fractional Snow Cover Dataset for China (2000-2022)

Jinliang Hou[1], Mingkai Zhang[1,2], Xiaohua Hao[1], Jifu Guo[3], Peng Dou[1], Ying Zhang[1*], Chunlin Huang[1,4*]

[1] 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
[2] University of Chinese Academy of Sciences, Beijing, 100094, China
[3] College of Information Science and Technology, Gansu Agricultural University, Lanzhou 730070, China
[4] Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China

Correspondence to: Ying Zhang (zhang_y@lzb.ac.cn), Chunlin Huang (huangcl@lzb.ac.cn)

Abstract. We present ChinaAI-FSC, the first large-scale, standardized, AI-ready fractional snow cover (FSC) sample collection for China, spanning 22 snow seasons from 2000 to 2022 and addressing a critical gap in long-term snow monitoring. The dataset consists of 47,728 samples (each 128×128 MODIS-pixel tiles), where high-resolution Landsat-5/7/8/9 and Sentinel-2 imagery provide consistent FSC reference labels. A total of 20 feature variables, including MODIS surface reflectance (bands 1-7), topographic attributes, forest and land cover information, and geolocation factors, were extracted to enable both point-scale and tile-scale spatially contextualized AI modelling. A structured and transparent workflow, encompassing systematic sample preparation, rigorous quality control, spatiotemporal sample partitioning, and standardized metadata, ensures reproducibility, physical consistency, and interoperability across machine learning and deep learning applications. Dataset reliability and AI-readiness were systematically evaluated using a novel "Four Layers-Four Domains-Fifteen Attributes (4L-4D-15A)" assessment protocol, covering data, information, system, and application dimensions. The quality, reliability, and usability of ChinaAI-FSC were demonstrated through three representative use cases: (1) benchmarking of six ML/DL models (ANN, SVR, RF, CNN, UNet, and ResNet), (2) validation of the standard MODIS FSC product, and (3) nationwide seamless FSC mapping. By providing harmonized, validated, and well-documented samples, ChinaAI-FSC establishes a unified foundation for AI-driven snow cover mapping, long-term monitoring, and cryosphere–hydrological modelling, promoting reproducible, interoperable, and next-generation research in cryospheric science. The dataset is publicly available from the National Tibetan Plateau Data Center (TPDC) at https://doi.org/10.11888/Cryos.tpdc.303034 (also accessible via https://cstr.cn/18406.11.Cryos.tpdc.303034) and from Zenodo at https://doi.org/10.5281/zenodo.17707386.

Key words: Fractional Snow Cover (FSC); AI-Ready; ML/DL; MODIS

1 Background and Motivation

Fractional Snow Cover (FSC) is a fundamental indicator for monitoring snowpack dynamics, as it quantifies the proportion of snow within a pixel, providing a continuous measure of snow extent that goes beyond binary snow/no-snow classifications.

Scientifically, FSC is a critical variable linking snowpack dynamics with energy and water exchanges at the land-atmosphere interface. It strongly influences surface albedo and shortwave radiation absorption, thereby affecting the timing of snowmelt, soil moisture evolution, and local to regional atmospheric circulation (Hall & Riggs, 2007; Frei et al., 2012). Accurate FSC information enhances the representation of snow processes in land surface and climate models, helping to reduce biases in surface energy budgets and improve projections of climate feedback (Thackeray & Fletcher, 2016; Mudryk et al., 2020). From an applied perspective, FSC is indispensable for hydrological forecasting and water resource management, as it governs meltwater contributions to rivers and reservoirs, influencing agricultural planning, flood risk assessment, and hydropower operations (Barnett et al., 2005). Moreover, Furthermore, long-term, high-accuracy FSC records are essential for detecting cryospheric responses to climate change, guiding adaptation strategies, and supporting international climate assessments such as those conducted by the IPCC. Despite its importance, operational estimation of FSC continues to face challenges arising from cloud contamination, complex terrain, vegetation canopy effects, and sensor limitations (Salomonson & Appel, 2004; Stillinger et al., 2023).

Early approaches to FSC retrieval primarily relied on statistical regression techniques, such as linear and exponential models, as well as spectral mixture analysis (Salomonson & Appel, 2004, 2006; Painter et al., 2009). Statistical regression models establish empirical relationships between FSC and a limited set of spectral band combinations. Although computationally straightforward, their performance is highly sensitive to regional and seasonal variability, often resulting in poor generalization and systematic biases under heterogeneous surface conditions (Hall et al., 2002; Raleigh et al., 2013; Xin et al., 2012). Spectral mixture models, in contrast, assume that pixel reflectance is a linear or nonlinear combination of pure endmembers, which can partially alleviate mixed-pixel effects. However, their accuracy heavily depends on the quality and representativeness of the endmember library, while the endmembers themselves can vary substantially with snow conditions and land-cover types. These limitations constrain the applicability of spectral mixture analysis in complex terrain and spatially heterogeneous environments (Dozier et al., 2008; Metsämäki et al., 2012; Painter et al., 2003, 2009; Rittger et al., 2013). Recent advances, such as the Multiple Endmember Spectral Mixture Analysis with Automated Global Endmember selection (MESMA-AGE), have partially alleviated these limitations by dynamically selecting optimal endmember combinations from large spectral libraries and accounting for variability in snow, vegetation, soil, and illumination conditions. This strategy has enabled improved sub-pixel FSC estimation over complex mountain environments, and has been successfully applied to generate daily MODIS FSC products for the Asian Water Tower region, which is characterized by extreme terrain, heterogeneous land cover, and highly variable snow conditions (Pan et al., 2024).

In recent years, FSC estimation has evolved from traditional empirical regression and spectral mixture decomposition methods toward data-driven machine learning (ML) and deep learning (DL) paradigms. Early studies applied artificial neural networks (ANNs) that integrated MODIS surface reflectance with auxiliary variables to improve FSC estimation (Dobreva & Klein, 2011; Hou & Huang, 2014). In challenging environments such as mountainous or forested regions, FSC retrieval has been addressed using algorithmic strategies that explicitly account for vegetation and terrain effects, as demonstrated by Czyzowska-Wisniewski et al. (2015) with ANNs and by Xiao et al. (2022) with ensemble tree-based models. Kuter et al. (2018, 2021,

2022) compared multiple ML algorithms, including multivariate adaptive regression splines (MARS), random forest (RF), and support vector regression (SVR), revealing the respective strengths of these traditional ML approaches for FSC modelling. With advances in big data and high-performance computing, research on FSC mapping has increasingly shifted toward DL-based multisource spatiotemporal fusion frameworks, especially in complex plateau and mountainous regions where snow distributions exhibit pronounced spatiotemporal heterogeneity. In this context, Azizi et al. (2024) and Liu et al. (2024) developed CNN-based architectures that explicitly exploit spatial texture and contextual information to characterize heterogeneous snow patterns in complex mountainous terrain. Similarly, Zhao et al. (2024) embedded radiative transfer constraints into DL frameworks, enhancing the physical consistency, accuracy, and robustness of FSC estimation. These advancements underscore the growing potential of combining DL with physically informed modelling to improve snow cover estimation in complex environments. Overall, ML- and DL-based approaches have markedly advanced FSC retrieval by capturing high-dimensional nonlinear relationships among spectral, topographic, and auxiliary variables, thereby substantially enhancing estimation accuracy and robustness.

Despite rapid progress in AI-driven FSC modelling, current studies still rely heavily on localized observations or limited experimental samples with narrow spatial coverage and short temporal spans. Although such models often achieve high accuracy in controlled settings, they struggle to scale effectively for large-area, long-term, and cross-regional AI-based FSC estimation. This limitation primarily stems from two interrelated challenges: (1) *Absence of large-scale FSC datasets suitable for AI-based modelling.* To date, no publicly available benchmark datasets exist that can be directly used to train, validate, and transfer AI-based FSC models across regions and snow regimes, severely constraining the generalization and applicability of current models. (2) *Lack of standardized protocols for FSC dataset construction and evaluation.* Currently, there is no unified framework for generating reference labels, selecting features, performing quality control, or evaluating dataset quality in a manner suitable for AI applications. This lack of standardized procedures prevents reproducibility, fair algorithm comparison, and transparent benchmarking. Collectively, these challenges highlight an urgent need to develop FSC datasets that are large-scale, standardized, and directly usable for AI-based modelling, providing a robust foundation for scalable snow monitoring and paving the way for the formal concept of AI-ready datasets introduced below.

The concept of *AI-ready data* refers to high-quality, standardized, and interoperable datasets explicitly optimized for AI applications, i.e., data that are immediately usable for model training, validation, and deployment without extensive preprocessing (Kidwai-Khan et al., 2024; Poduval et al., 2023). AI-ready datasets are characterized by *end-to-end curation*, encompassing data acquisition, cleaning, calibration, quality control, and metadata standardization to ensure traceability, reproducibility, and interoperability. The U.S. National Science Foundation's *National AI Research Resource (NAIRR) Pilot* exemplifies this paradigm, fostering open, cross-domain sharing of AI-ready data to advance education, research, and model development (NSF, 2024). These initiatives mark a broader global shift from traditional data archiving toward AI-oriented data infrastructures designed to directly empower model-driven science. In the context of FSC mapping, constructing an AI-Ready FSC dataset entails systematically and intelligently organizing multi-source, heterogeneous snow-related environmental variables together with FSC reference labels into a cohesive, structured, and AI-oriented framework. Such a dataset enables

direct AI model training without extensive preprocessing, facilitates reproducible and scalable FSC estimation, and supports transparent benchmarking and cross-regional generalization.

Building upon this concept, an AI-ready FSC dataset can be envisioned as a large-scale, standardized collection of snow cover samples that adhere to the following key principles: (1) Spatiotemporal representativeness. Samples should comprehensively capture diverse terrains, land-cover types, and snow conditions across multiple temporal scales. (2) Physical and environmental completeness. The dataset should incorporate well-defined environmental and geophysical variables that govern snow accumulation, melting, and redistribution processes. (3) High-quality reference FSC labelling. Reliable FSC reference labels should be derived from multi-source, high-resolution remote sensing observations, with rigorous spatial and temporal consistency checks. (4) Standardized AI-ready metadata and valuation protocols. Comprehensive metadata and unified standards for dataset construction, documentation, and performance evaluation are essential to ensure reproducibility and comparability across studies. By adhering to these principles, an AI-Ready FSC dataset provides a robust foundation for advancing intelligent, physically consistent, and scalable FSC modelling, bridging the current gap between algorithmic innovation and data standardization.

This study presents ChinaAI-FSC, a standardized and AI-ready MODIS FSC dataset for China spanning 22 snow seasons from 2000 to 2022. Rather than developing new FSC retrieval algorithms, this work focuses on the methodological challenges of constructing, validating, and evaluating FSC datasets for large-scale ML and DL applications. The study systematically describes the study area and multisource data used for reference FSC generation and feature construction, and details a reproducible dataset construction pipeline, including feature-FSC matching, consistency-based quality control, and sample quality assessment. Large-scale benchmark and validation experiments across China are further conducted to demonstrate the reliability and representativeness of the dataset under diverse climatic and topographic conditions. By constructing and applying ChinaAI-FSC within this framework, the study establishes a methodological innovation for AI-ready FSC dataset development at continental scale. Specifically, it introduces a unified AI-Ready FSC sample repository, a transparent and reproducible construction workflow, and a formal Four Layers-Four Domains-Fifteen Attributes (4L-4D-15A) AI-readiness evaluation framework, providing a standardized methodological paradigm for AI-driven snow monitoring.

2 Study Area

This study focuses on the entire terrestrial domain of China, which exhibits remarkable representativeness and diversity in the global snow cover distribution. China's complex topography and diverse climate regimes give rise to three major stable snow regions: the Qinghai-Tibetan Plateau, the Northeast-Inner Mongolia, and Northern Xinjiang (Tan et al., 2019). The Qinghai-Tibetan Plateau is characterized by high elevations and alpine conditions, where snow dynamics are jointly influenced by monsoon and radiative processes. The Northeast-Inner Mongolia region, dominated by forested landscapes, displays strong snow-vegetation interactions typical of mid- to high-latitude ecosystems. In contrast, Northern Xinjiang represents a continental mountain snow environment under the westerlies, exhibiting pronounced vertical stratification and high interannual

variability. Beyond these major snow regions, other parts of China, such as the Loess Plateau, the North China Plain, and the southwestern mountains, experience transient and intermittent snow events during winter. Overall, China encompasses highly diverse and contrasting snow regimes, ranging from persistent multi-year snow zones to short-lived seasonal snow cover. This pronounced spatial and temporal heterogeneity provides an ideal foundation for constructing a comprehensive and representative AI-ready FSC dataset capable of supporting large-scale and multi-scenario modelling. Accordingly, the study domain is divided into four subregions: Xinjiang (XJ), Qinghai-Tibetan Plateau (QTP), Northeast-Inner Mongolia (NE-IM), and Other regions (Fig. 1).

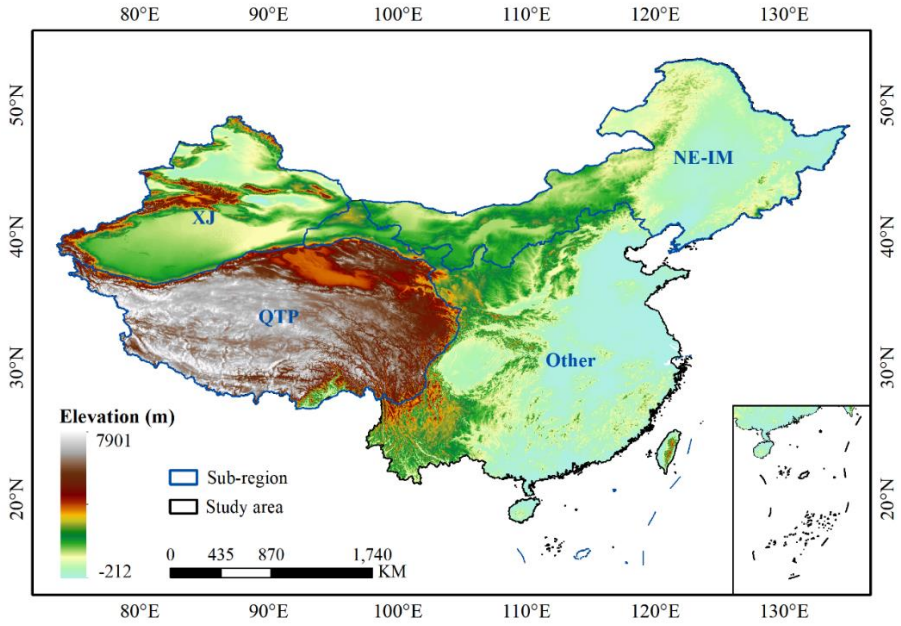

Figure 1: Overview of the study area. The entire study domain is divided into four subregions: Xinjiang (XJ), Qinghai-Tibetan Plateau (QTP), Northeast-Inner Mongolia (NE-IM), and Other regions.

3 Development of the AI-Ready MODIS FSC Dataset

This study aims to construct the first standardized, AI-ready MODIS FSC sample database for China (ChinaAI-FSC), spanning 22 snow seasons (from October to March of the following year) from 2000 to 2022. The detailed construction workflow is illustrated in Fig. 2.

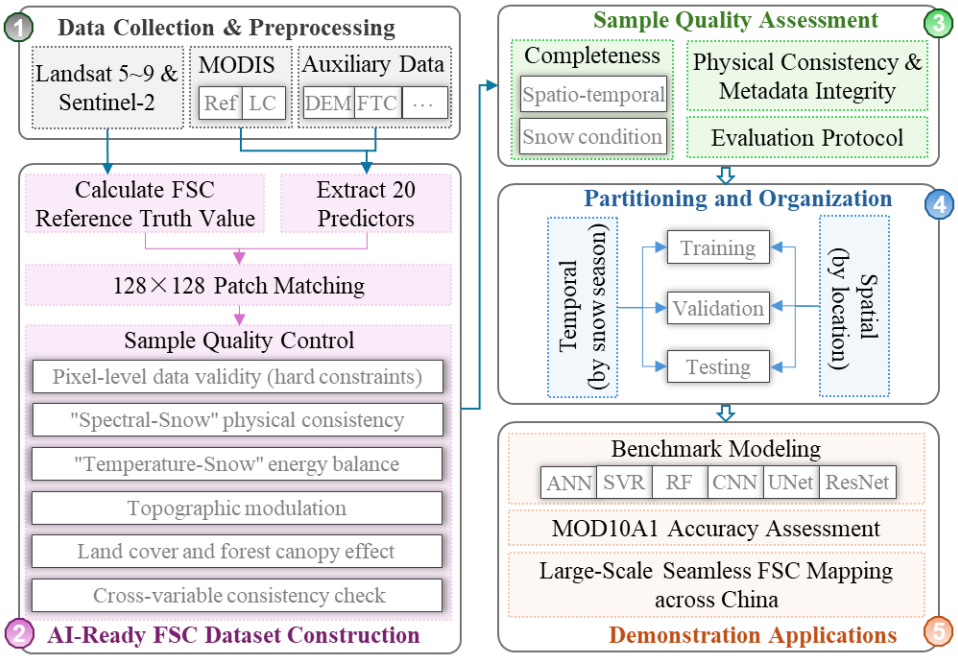

Figure 2: Workflow for constructing the AI-ready MODIS FSC sample dataset. Ref denotes surface reflectance, LC denotes land cover type, FTC denotes fractional tree cover, and MOD10A1 denotes the standard MODIS FSC product. ANN, SVR, RF, CNN, UNet, and ResNet denote Artificial Neural Network, Support Vector Regression, Random Forest, Convolutional Neural Network, U-shaped Convolutional Neural Network, and Deep Residual Network, respectively.

3.1 Data Collection & Preprocessing

3.1.1 Landsat and Sentinel-2 imagery

The Landsat Collection 2 Level-2 Surface Reflectance (SR) products for Landsat-5 TM, Landsat-7 ETM+, Landsat-8 OLI, and Landsat-9 OLI-2 were obtained from the U.S. Geological Survey (USGS) Earth Explorer platform (https://earthexplorer.usgs.gov/). Sentinel-2A/2B MSI Level-2A SR products were sourced from the European Space Agency (ESA) Copernicus Open Access Hub (https://scihub.copernicus.eu/). These products provide atmospherically corrected surface reflectance suitable for quantitative analysis (Masek et al., 2006; Vermote et al., 2016; Louis et al., 2016). To ensure data quality, only scenes with an overall cloud cover below 15% were selected across 22 snow seasons between 2000 and 2022. The number of raw scenes collected for each subregion and snow season is summarized in Table 1.

Systematic preprocessing, including quality screening, spatial reconstruction, and radiometric harmonization, was applied across all sensors. For Landsat imagery, residual low-quality pixels, including clouds, cirrus, and cloud shadows, were identified using the QA_PIXEL band generated by the CFMask algorithm (Zhu et al., 2015), while for Sentinel-2 imagery, cloud and shadow pixels were masked using the Scene Classification Layer (SCL) (Drusch et al., 2012). Low-quality or masked pixels in both datasets were subsequently reconstructed using a spectrally constrained spatial gap-filling approach, in which missing reflectance values were estimated from neighbouring valid pixels with similar spectral characteristics to preserve local

spectral consistency and spatial continuity (Chen et al., 2011). For Landsat-7 ETM+ images acquired after the Scan Line Corrector (SLC) failure in May 2003, scan-line gaps were corrected using a local neighbourhood linear interpolation, where missing pixels were replaced by values derived from adjacent valid observations along the scan-line direction and smoothed with a 3×3 spatial kernel (Markham et al., 2004; Chen et al., 2011). Additionally, due to orbit drift and the consequent degradation of observation quality after 2021 (Qiu et al., 2021), Landsat-7 ETM+ data acquired from 2021 onwards were excluded. Optionally, topographic correction was applied in mountainous regions using the C-correction method to reduce illumination-induced reflectance variability (Crawford et al., 2023). To address radiometric discrepancies among sensors, cross-sensor normalization was performed using Landsat-9 OLI-2 as the reference, with regression-based adjustments applied to overlapping scenes to harmonize Landsat-8/7/5 and Sentinel-2 MSI surface reflectance (Chander et al., 2013; Claverie et al., 2018). Finally, multi-sensor mosaicking was conducted for images acquired on the same day, and the mosaicked imagery was reprojected to a geographic coordinate system with a uniform spatial resolution of 0.00833° (~30 m).

Table 1: Number of Landsat and Sentinel-2 Images Used for 2000-2022 Snow Seasons

Snow Season	XJ					NE-IM				
	LT05	LE07	LC08	LC09	S2	LT05	LE07	LC08	LC09	S2
2000-2001	127	265	-	-	-	342	565	-	-	-
2001-2002	169	222	-	-	-	446	436	-	-	-
2002-2003	143	280	-	-	-	474	648	-	-	-
2003-2004	107	215	-	-	-	414	131	-	-	-
2004-2005	142	235	-	-	-	509	64	-	-	-
2005-2006	74	273	-	-	-	264	62	-	-	-
2006-2007	143	249	-	-	-	569	83	-	-	-
2007-2008	19	280	-	-	-	61	305	-	-	-
2008-2009	184	306	-	-	-	202	175	-	-	-
2009-2010	275	274	-	-	-	190	260	-	-	-
2010-2011	331	313	-	-	-	260	462	-	-	-
2011-2012	29	376	-	-	-	7	452	-	-	-
2012-2013	-	306	-	-	-	-	350	-	-	-
2013-2014	-	325	367	-	-	-	438	474	-	-
2014-2015	-	353	408	-	-	-	483	522	-	-
2015-2016	-	427	316	-	-	-	501	629	-	-
2016-2017	-	318	272	-	-	-	538	567	-	-
2017-2018	-	385	405	-	42	-	476	393	-	63
2018-2019	-	380	433	-	1160	-	276	369	-	1219
2019-2020	-	409	416	-	2089	-	329	586	-	3749
2020-2021	-	184	299	-	1951	-	121	647	-	2558
2021-2022	-	-	421	189	2595	-	-	590	419	1914
Total	**1743**	**6375**	**3337**	**189**	**7837**	**3738**	**7155**	**4777**	**419**	**9503**

Table 1: (continued)

Snow Season	QTP					Other					China
	LT05	LE07	LC08	LC09	S2	LT05	LE07	LC08	LC09	S2	
2000-2001	466	491	-	-	-	146	109	-	-	-	2511
2001-2002	424	488	-	-	-	143	103	-	-	-	2431
2002-2003	324	561	-	-	-	158	204	-	-	-	2792
2003-2004	396	345	-	-	-	131	201	-	-	-	1940
2004-2005	449	340	-	-	-	174	193	-	-	-	2106
2005-2006	227	574	-	-	-	70	175	-	-	-	1719
2006-2007	483	376	-	-	-	120	170	-	-	-	2193

2007-2008	155	468	-	-	-	35	175	-	-	-	1498
2008-2009	606	630	-	-	-	90	121	-	-	-	2314
2009-2010	539	576	-	-	-	113	138	-	-	-	2365
2010-2011	606	564	-	-	-	119	113	-	-	-	2768
2011-2012	97	655	-	-	-	15	116	-	-	-	1747
2012-2013	-	695	-	-	-	-	183	-	-	-	1534
2013-2014	-	721	767	-	-	-	160	98	-	-	3350
2014-2015	-	677	540	-	-	-	103	86	-	-	3172
2015-2016	-	676	274	-	-	-	105	123	-	-	3051
2016-2017	-	645	702	-	-	-	101	95	-	-	3238
2017-2018	-	237	747	-	166	-	141	131	-	25	3211
2018-2019	-	310	698	-	628	-	113	102	-	174	5862
2019-2020	-	340	271	-	1320	-	182	122	-	467	10280
2020-2021	-	219	521	-	1255	-	88	134	-	400	8377
2021-2022	-	-	670	338	4201	-	-	102	63	542	12044
Total	**4772**	**10588**	**5190**	**338**	**7570**	**1314**	**2994**	**993**	**63**	**1608**	**80503**

Note: LT05, LE07, LC08, LC09, and S2 denote Landsat-5 TM, Landsat-7 ETM+, Landsat-8 OLI, Landsat-9 OLI-2, and Sentinel-2, respectively

3.1.2 MODIS data

The MODIS series products utilized in this study include surface reflectance bands 1-7, the standard MODIS snow product (MOD10A1, Collection 6), and the MODIS land cover dataset (MCD12Q1, Collection 6). The surface reflectance data was obtained from the Global 500 m seamless dataset of MODIS-derived land surface reflectance (SDC500) for 2000-2022 (Liang et al., 2024). Unlike the standard MOD09GA, SDC500 generates a continuous daily 500 m reflectance time series by correcting BRDF effects, detecting outliers, and filling missing values with phenology-guided spline interpolation. Snow and snow-free periods are treated separately to preserve seasonal reflectance dynamics. The dataset achieves high accuracy, with a mean absolute error of only 0.043, providing a reliable basis for FSC retrieval. All MODIS datasets were subsequently reprojected and resampled to a common geographic coordinate system with a spatial resolution of 0.005° (~500 m).

3.1.3 Auxiliary Data

Additional environmental factors influencing snow distribution were incorporated, including forest cover, land surface temperature (LST), and topographic variables derived from digital elevation data (DEM). Forest cover was represented using the global annual fractional tree cover dataset for 2000-2021 at 250 m resolution (GLOBMAP FTC), which accurately captures both global and regional forest dynamics (Liu et al., 2024). LST data were obtained from a daily 1 km all-weather land surface temperature dataset over China and surrounding regions (TRIMS LST), which shows high agreement with MODIS LST products in both magnitude and spatial distribution, with mean daytime and nighttime biases of 0.09 K and -0.03 K, and standard deviations of 1.45 K and 1.17 K, respectively (Tang et al., 2024). Topographic information was derived from the Shuttle Radar Topography Mission (SRTM) Version 4.1 DEM, accessed via the CGIAR-CSI database (Jarvis et al., 2008). Considering the pronounced seasonal periodicity of snow cover, Julian day was also included as an auxiliary variable. All auxiliary datasets were resampled to match the spatial resolution and coordinate system of the MODIS land surface products.

In addition, snow depth observations from a total of 507 meteorological stations distributed across mainland China (Fig. 1) were collected for seven snow seasons (2013-2020) and used for independent validation of the MODIS reference FSC.

3.2 AI-Ready FSC Dataset Construction

3.2.1 Calculation of MODIS Reference FSC

The standard SNOMAP algorithm (Hall et al., 1995) derives binary snow cover from the Normalized Difference Snow Index (NDSI), calculated using visible and shortwave infrared reflectance. However, in forested regions, snow detection accuracy can be substantially reduced by canopy occlusion and mixed-pixel effects, resulting in underestimation and omission errors. To address these limitations, Klein et al. (1998) proposed an improved SNOMAP algorithm that integrates a canopy reflectance model into the original NDSI framework. This modification explicitly separates the reflectance contributions of the canopy and the underlying snow surface through canopy reflectance and transmittance, thereby improving snow detection accuracy in vegetated areas. In this study, the improved SNOMAP algorithm was applied to the pre-processed Landsat and Sentinel-2 imagery to generate binary snow maps at 30 m spatial resolution. Based on these high-resolution snow maps, FSC reference values at the MODIS scale were estimated. For each MODIS pixel, a circular neighbourhood was centred on the MODIS pixel centroid with a radius equal to 1.5 times the MODIS pixel size, to account for potential geolocation discrepancies between MODIS and Landsat/Sentinel-2 imagery. The MODIS FSC reference value was then computed as the proportion of snow-covered Landsat/Sentinel-2 pixels within this neighbourhood (Dobreva and Klein, 2011).

3.2.2 Extraction of Feature Variables

A total of twenty feature variables were derived for FSC modelling (Table 2), including MODIS surface reflectance bands 1-7 (denoted as Ref1-Ref7) from the SDC500 product, NDSI, Normalized Difference Vegetation Index (NDVI), land cover (LC), LST, FTC, and topographic factors. NDSI and NDVI were computed from the SDC500 reflectance using bands 4 and 6, and bands 2 and 1, respectively. Five topographic variables, i.e., elevation, slope, aspect, terrain relief, and surface roughness, were derived from the DEM.

Table 2: Description of the 20 Input Features Integrated in the AI-Ready FSC Dataset

Variable	Data Product	Data Source	Reference
Ref1-Ref7	SDC500	https://data-starcloud.pcl.ac.cn/iearthdata/27	Liang et al. (2024)
NDSI			-
NDVI			-
LC	MCD12Q1	https://lpdaac.usgs.gov/products/mcd12q1v061/	-
LST	TRIMS LST	https://data.tpdc.ac.cn/zh-hans/data/05d6e569-6d4b-43c0-96aa-5584484259f0	Tang et al. (2024)
FTC	GLOBMAP FTC	https://zenodo.org/records/10589730	Liu et al. (2024)
Elevatio, Slope, Aspect, Terrain Relief, Surface Roughness	SRTM DEM	https://srtm.csi.cgiar.org/srtmdata/	Jarvis et al. (2008)
Longitude, Latitude	-	-	-

Julian Day	-	-	-

3.2.3 Generation of the Original FSC Sample Dataset

Considering the distinct structural requirements of different ML and DL models, the sample was organized to accommodate both point-based and spatially continuous inputs. Traditional models such as ANNs and SVR typically operate on discrete point samples, whereas convolution-based models (e.g., CNNs) require spatially continuous image blocks. Considering the spatial resolution of MODIS data and the spatial coverage of individual high-resolution Landsat 5/7/8/9 and Sentinel-2 scenes, the study area was divided into regular 0.64° × 0.64° grid tiles, each corresponding to 128 × 128 MODIS pixels. Each tile was assigned a unique row-column identifier based on its geographic position (Fig. 3). This spatial partitioning provides an optimal balance between areal coverage and computational efficiency while maintaining high decomposability. It also supports flexible aggregation into multi-scale tiles (e.g., 8×8, 16×16, 32×32, and 64×64 MODIS pixels), enabling adaptive feature extraction and model training across different spatial scales. Within each grid tile, twenty feature variables (input features) were extracted from MODIS and auxiliary datasets and paired with FSC reference values derived from Landsat/Sentinel-2 observations based on precise spatiotemporal correspondence. These "feature-reference FSC" matchups collectively form the original MODIS FSC sample dataset. In addition, pixels with FSC ≥ 15% (Painter et al., 2009; Zhang et al., 2019) were defined as snow-covered, and samples with mean snow-covered fractions <5% or >95% were excluded, corresponding to nearly snow-free and fully snow-covered homogeneous conditions. Such samples contain little internal variability and provide limited value for learning fractional snow-land relationships. Consequently, ChinaAI-FSC contains no completely snow-free or fully snow-covered samples, but focuses on spatially heterogeneous mixed snow conditions. Over the 22 snow seasons, a total of 166,763 original samples (each comprising 128×128 MODIS pixels) were generated (Table 3), and each sample is represented as a spatial tile with a corresponding reference FSC target, and co-registered multi-source features describing terrain, surface properties, and environmental conditions, including longitude, latitude, elevation, aspect, slope, terrain relief, surface roughness, LC, LST, FTC, DOY, NDVI, NDSI, and seven reflectance bands (Ref1-Ref7), respectively.

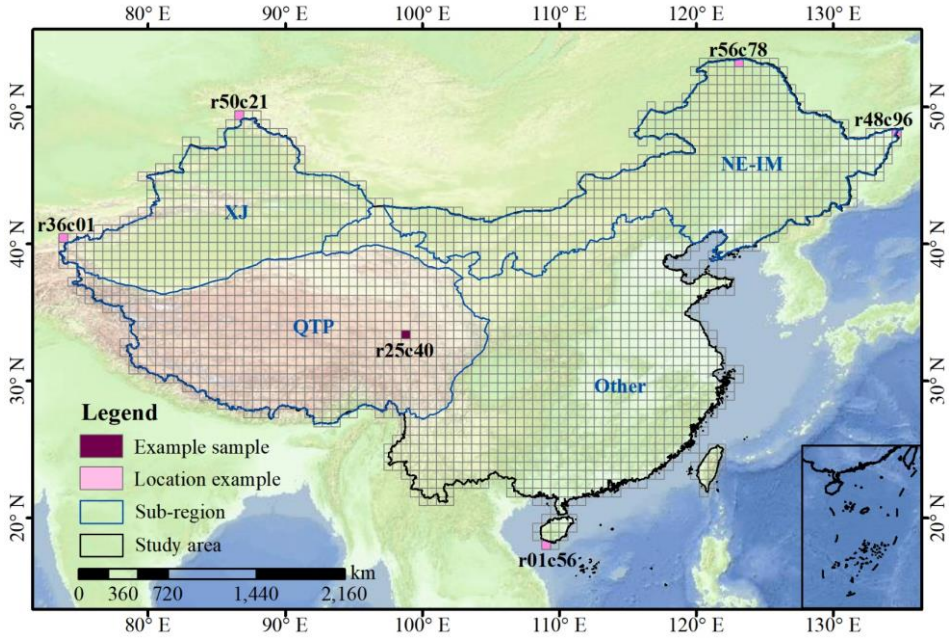

Figure 3: Spatial partitioning of the study area into regular 0.64° × 0.64° grid tiles, each corresponding to 128 × 128 MODIS pixels. Tiles are assigned unique row and column identifiers based on spatial location, with column numbers (c01, c02, …) increasing from west to east and row numbers (r01, r02, …) increasing from south to north, as illustrated by the pink "Location example". The marked example sample indicates the location of the representative case discussed in Section 3.3.2.

Table 3: Statistics of Original FSC Samples over 22 Snow Seasons

Snow Season	XJ	NE-IM	QTP	Other	China
2000-2001	1107	3142	3038	395	7682
2001-2002	1047	2892	2899	343	7181
2002-2003	1209	3477	2856	652	8194
2003-2004	1076	1608	2319	440	5443
2004-2005	1208	1846	2516	609	6179
2005-2006	1146	968	2498	344	4956
2006-2007	1269	2172	2725	424	6590
2007-2008	1064	1112	1992	398	4566
2008-2009	1324	1136	3912	256	6628
2009-2010	1375	1453	3612	394	6834
2010-2011	1741	2417	3730	302	8190
2011-2012	1162	1457	2380	202	5201
2012-2013	889	1117	2260	296	4562
2013-2014	1618	2975	4688	388	9669
2014-2015	1817	3185	3690	223	8915
2015-2016	1737	3298	2836	356	8227
2016-2017	1389	3568	4053	237	9247
2017-2018	1807	2717	2848	441	7813
2018-2019	2518	2286	2853	327	7984
2019-2020	3006	5339	1966	653	10964
2020-2021	2504	4222	2429	485	9640
2021-2022	3257	4535	3726	580	12098
Total	**35270**	**56922**	**65826**	**8745**	**166763**

3.2.4 Sample Quality Control

To guarantee the physical consistency and high reliability of the FSC sample dataset, a comprehensive quality control (QC) was implemented. The QC procedures were applied at two hierarchical levels, pixel and tile, integrating physically motivated constraints with multi-level consistency checks. These procedures collectively ensured that the resulting dataset is physically coherent, statistically robust, and fully AI-ready for subsequent model training and evaluation (Fig. 4).

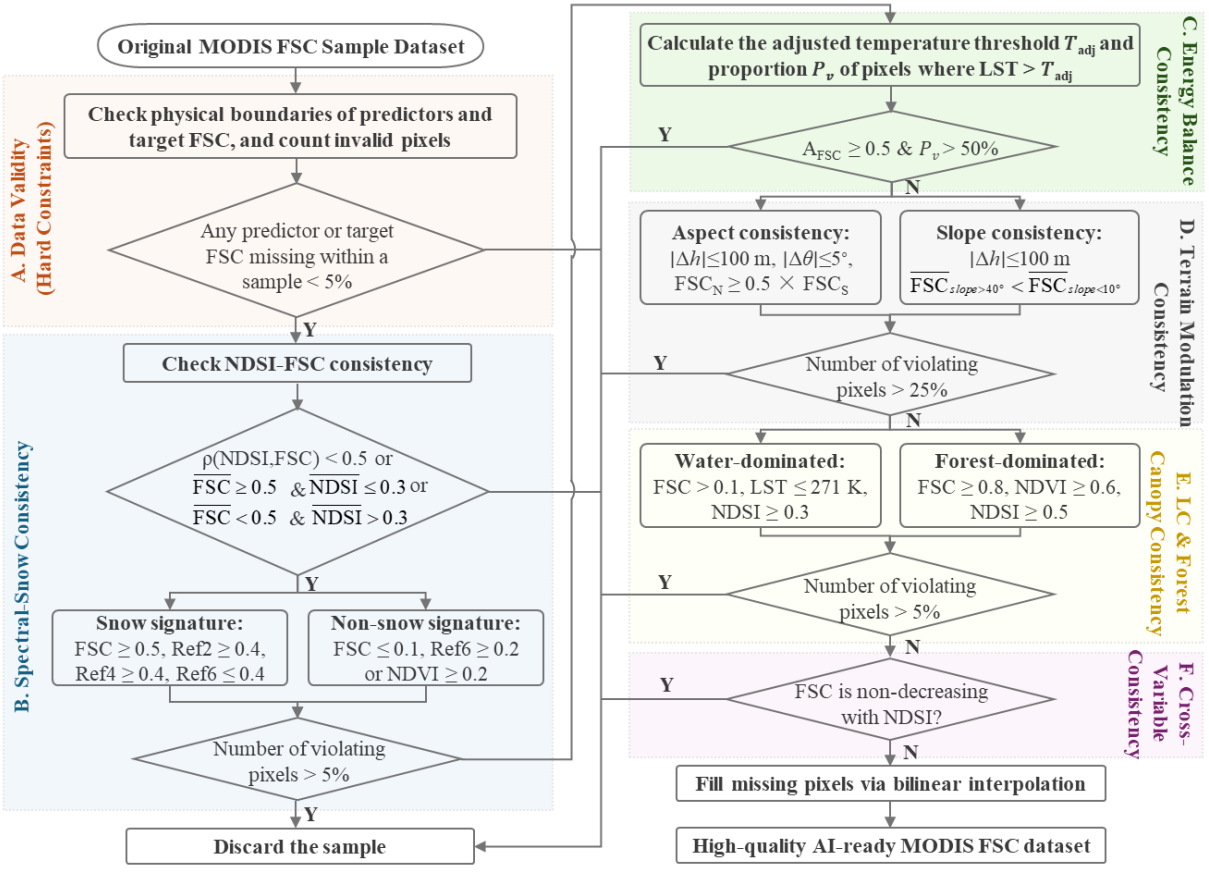

Figure 4: Technical Workflow of FSC Sample Quality Control.

A. Pixel-level data validity (hard constraints)

Pixel-level validity checks were conducted to eliminate physically inconsistent or anomalous samples. Specifically, physical range constraints were imposed on all feature variables and the target FSC variable according to established physical limits. For instance, surface reflectance for all bands was required to satisfy $0 \leq \rho \leq 1$; NDVI ∈ [-1, 1]; NDSI ∈ [-1, 1]; LST ∈ [180 K, 340 K]; FTC ∈ [0, 1]; FSC ∈ [0, 1]; and slope ∈ [0°, 60°], where slopes > 60° typically indicate noise or projection artifacts. Pixels exceeding these limits were flagged as invalid. Additionally, for each sample tile, if the proportion of missing pixels in

any variable exceeded 5%, the sample was discarded to ensure data validity and statistical reliability. Thresholds were chosen conservatively to remove only clearly invalid pixels.

B. Spectral-Snow Physical Consistency

To ensure physical consistency between spectral characteristics and snow cover, two complementary explicit constraints were considered: NDSI-FSC relationship verification and spectral signature constraint checking. Here, it should be noted that the relevant thresholds were set based on known snow spectral properties and prior studies (Dozier, 1989; Hall et al., 2002), and were chosen conservatively to retain representative tiles while excluding extreme anomalies.

① **NDSI-FSC Consistency:** For each sample tile, the consistency between NDSI and FSC was quantified using the Pearson correlation coefficient computed across all pixels within the tile (Eq. (1)):

$$\rho(\mathrm{NDSI},\ \mathrm{FSC}) = \frac{\sum_{i=1}^{128}\sum_{j=1}^{128}\left(\mathrm{NDSI}_{ij} - \overline{\mathrm{NDSI}}\right)\left(\mathrm{FSC}_{ij} - \overline{\mathrm{FSC}}\right)}{\sqrt{\sum_{i=1}^{128}\sum_{j=1}^{128}\left(\mathrm{NDSI}_{ij} - \overline{\mathrm{NDSI}}\right)^2}\sqrt{\sum_{i=1}^{128}\sum_{j=1}^{128}\left(\mathrm{FSC}_{ij} - \overline{\mathrm{FSC}}\right)^2}} \tag{1}$$

Here, NDSI_{ij} and FSC_{ij} denote the NDSI and FSC values at pixel (i, j) within the tile, and $\overline{\mathrm{NDSI}}$ and $\overline{\mathrm{FSC}}$ are the corresponding tile-averaged values. Samples with $\rho(\mathrm{NDSI},\ \mathrm{FSC}) < 0.5$ was considered spectrally inconsistent with snow physics and were excluded. In addition, tile-level constraints were imposed to avoid contradictory snow signals: samples with $\overline{\mathrm{FSC}} \geq 0.5$ and $\overline{\mathrm{NDSI}} < 0.3$, or $\overline{\mathrm{FSC}} < 0.5$ and $\overline{\mathrm{NDSI}} \geq 0.3$, were also removed.

② **Spectral Signatures of Snow Pixels:** For pixels with FSC ≥ 0.5, spectral reflectance was required to satisfy the following physical constraints: green band reflectance Ref4 ≥ 0.2, near-infrared reflectance Ref2 ≥ 0.4, and shortwave infrared reflectance Ref6 ≤ 0.4. If more than 5% of pixels in a sample failed to meet these conditions, the sample was excluded.

③ **Spectral Signatures of Non-Snow Pixels:** For pixels with FSC ≤ 0.1, the conditions Ref6 ≥ 0.2 or NDVI ≥ 0.2 must both be met. Samples were discarded if more than 5% of pixels failed to meet these constraints.

C. Temperature-Snow Energy Balance Consistency

To assess the consistency of energy balance, the relationship between LST and snow cover was evaluated. An elevation-adjusted cooling constraint was applied to account for topographic effects on temperature, assuming that a 1000 m increase in elevation corresponds to a 2K relaxation of the threshold. The pixel-wise adjusted temperature threshold T_{adj} was calculated using Eq. (2), with elevation aligned to the LST data. For each sample tile, the proportion of pixels exceeding this threshold, denoted as P_v, was computed. Samples with $\overline{\mathrm{FSC}} \geq 0.5$ and $P_v > 0.5$ were considered inconsistent with the energy balance after accounting for elevation effects and were consequently removed from the dataset.

$$T_{adj}(i,j) = T_{base} + 2K \times \frac{h(i,j)}{1000m} \tag{2}$$

Here, T_{base} is the base temperature threshold, set to 273.15 K, and $h(i,j)$ denotes the elevation of pixel (i,j). Thresholds were defined around the physical freezing point to remove only clearly inconsistent samples while retaining representative snow conditions.

D. Topographic Modulation Consistency

Given the strong influence of topography on snow distribution, additional consistency checks were applied based on aspect and slope:

① **Aspect Consistency**: Under comparable elevation (±100 m) and slope (±5°) conditions, north-facing (315°-45°) shaded slopes are generally expected to retain comparable or greater snow cover than south-facing (135°-225°) sun-exposed slopes. Accordingly, FSC on north-facing slopes was required to be not lower than 50% of that on south-facing slopes. Samples were excluded as topographically inconsistent if more than 25% of pixels violated this condition (Eq. (3)).

$$\begin{cases} |\Delta h| \le 100m \\ |\Delta \theta| \le 5^{\circ} \\ FSC_{\text{north}} \ge 0.5 \times FSC_{\text{south}} \end{cases} \tag{3}$$

②**Slope Consistency**: Under comparable elevation conditions, snow cover on steep slopes (>40°) is generally lower than that on gentle slopes (<10°) due to enhanced redistribution and melt. Samples in which more than 25% of pixels contradicted this expectation were excluded as slope-inconsistent (Eq. (4)).

$$\begin{cases} |\Delta h| \le 100m \\ \overline{FSC}_{\text{slope}>40^{\circ}} < \overline{FSC}_{\text{slope}<10^{\circ}} \end{cases} \tag{4}$$

Thresholds for aspect and slope were selected to reflect broad, well-established topographic tendencies rather than local-scale variability (Grünewald et al., 2013).

E. Land Cover and Forest Canopy Effect Consistency

①**Water Body Consistency Constraint:** Samples containing more than 60% water pixels are classified as water-dominated. For such samples, pixels with FSC > 0.1 were required to satisfy LST ≤ 271 K and NDSI ≥ 0.3, corresponding to plausible frozen or ice-covered conditions. These thresholds are informed by the spectral and thermal properties of ice and snow reported in prior studies (Dozier, 1989; Hall et al., 2002) and are set conservatively to retain representative ice-covered observations while excluding non-frozen water pixels that could be misclassified as snow. If more than 5% of pixels fail to meet this criterion, the sample is considered spectrally and thermally inconsistent and is excluded.

②**Forest Canopy Obstruction Constraint:** Samples with FTC ≥ 60% were classified as forest-dominated, where snow signals are prone to systematic underestimation due to canopy obstruction. Samples were excluded if more than 5% of pixels exhibited a combination of high snow cover (FSC ≥ 0.8) together with high NDVI (≥ 0.6) and high NDSI (≥ 0.5), which indicates a physically implausible snow signal under dense vegetation. Thresholds were guided by established canopy-snow interaction mechanisms (Essery and Pomeroy, 2004).

F. Cross-Variable Consistency Check

Finally, the internal consistency between FSC and NDSI is evaluated. Pixels within each sample were divided into 10 equal-width bins according to NDSI values, and the mean FSC was calculated for each bin. In theory, the mean FSC should increase monotonically with NDSI. Samples exhibiting two or more pronounced reversals, defined as a decrease in mean FSC exceeding 0.1, were regarded as anomalous and excluded.

Table 4 summarizes how the number of samples changed during the six quality control steps. After applying these six rigorous QC steps described above, the few remaining missing pixels in the samples were filled using bilinear interpolation. The resulting AI-ready, high-quality MODIS FSC sample dataset is summarized in Table 5.

Table 4: Number of original, excluded, and remaining samples after each QC step.

QC steps	XJ		NE-IM		QTP		Other		China	
	Exc	Rem	Exc	Rem	Exc	Rem	Exc	Rem	Exc	Rem
Original		35270		56922		65826		8745		166763
A	20454	14816	31462	25460	30394	35432	8010	735	90320	76443
B.①	1446	13370	4484	20976	2934	32498	69	666	8933	67510
B.②	1669	11701	2304	18672	354	32144	0	666	4327	63183
B.③	2692	9009	1525	17147	8220	23924	308	358	12745	50438
C	872	8137	9	17138	0	23924	0	358	881	49557
D.①	484	7653	208	16830	518	23406	1	357	1211	48246
D.②	0	7653	0	16830	0	23406	0	357	0	48246
E.①	0	7653	0	16830	0	23406	0	357	0	48246
E.②	0	7653	0	16830	0	23406	0	357	0	48246
F	93	7560	299	16531	124	23282	2	355	518	47728

Note: Exc and Rem indicate, respectively, the number of samples excluded and the number of samples remaining at each step.

Table 5: Statistics of AI-Ready MODIS FSC samples after quality control.

Snow Season	XJ	NE-IM	QTP	Other	China
2000-2001	267	1081	1116	9	2473
2001-2002	232	793	1007	9	2041
2002-2003	284	1012	1231	48	2575
2003-2004	165	436	786	22	1409
2004-2005	177	609	930	25	1741
2005-2006	197	325	837	7	1366
2006-2007	234	640	1025	8	1907
2007-2008	219	352	751	19	1341
2008-2009	313	326	1601	4	2244
2009-2010	290	313	1376	18	1997
2010-2011	387	656	1287	7	2337
2011-2012	250	423	836	9	1518
2012-2013	181	355	739	33	1308
2013-2014	314	1035	1660	8	3017
2014-2015	410	930	1256	1	2597
2015-2016	384	952	865	26	2227
2016-2017	229	1113	1195	7	2544
2017-2018	360	737	813	0	1910
2018-2019	466	612	1070	8	2156
2019-2020	747	1411	824	30	3012
2020-2021	595	1062	640	25	2322
2021-2022	859	1358	1437	32	3686
Total	**7560**	**16531**	**23282**	**355**	**47728**

3.3 Sample Quality Assessment

To ensure that the ChinaAI-FSC dataset is scientifically reliable for AI-based FSC modelling, we conducted a multi-dimensional sample quality assessment to determine whether the constructed feature-FSC matchups are sufficiently complete, physically consistent, and statistically representative to support AI learning across diverse snow regimes. We also introduce a novel set of AI-readiness evaluation protocols that explicitly assess learning validity, systematically quantifying the degree to which the dataset supports AI-driven modelling.

3.3.1 Sample Completeness

Sample completeness is a prerequisite for ensuring the learning validity of AI-based FSC models. The completeness of the AI-ready FSC dataset was evaluated in terms of spatial, temporal, and snow-cover representativeness to ensure the absence of systematic bias or tilt.

Spatial completeness was assessed by examining whether the samples provide sufficient coverage across the full geographic extent of China and its major snow regimes. Fig. 5a-b compare the spatial distributions of samples before and after QC. Although QC inevitably reduced the total number of samples, the final dataset preserves broad and balanced spatial coverage across the three dominant stable snow regions. No systematic regional gaps or clustering were introduced, indicating that the dataset maintains sufficient spatial representativeness for training and evaluating AI models at regional to continental scales.

Temporal completeness was evaluated by analysing the temporal distribution of samples across the 22 snow seasons. As summarized in Table 5, samples are consistently available across years and across all subregions, with no pronounced temporal discontinuities. This balanced temporal coverage ensures that interannual variability in snow conditions is sufficiently captured, providing a stable foundation for AI models aimed at long-term FSC monitoring and generalization across different snow seasons.

Snow-cover completeness was assessed at the tile-level by examining the distribution of sample mean FSC values across the range of 5% to 95%. It should be noted that at the pixel level, FSC values span the full 0-100% range, indicating that pixels with both low and high snow fractions are abundant within each tile. As shown in Fig. 6, the original dataset exhibits a strong skew toward low-FSC conditions, especially over the QTP, reflecting the predominance of patchy and subpixel snow in China. After QC, the FSC distribution became markedly more balanced, with improved representation of intermediate and high-FSC categories. This adjustment is critical for AI-based regression models, which require sufficient samples across the full FSC range to avoid biased learning.

To verify that the quality control procedures did not distort the underlying feature-FSC relationships, kernel density distributions of all 20 predictor variables were compared before and after screening (Fig. 7). The close agreement between the two distributions in terms of overall shape, central tendency, and tail behaviour indicates that the statistical structure of the original dataset is well preserved. This demonstrates that the final AI-ready FSC dataset achieves enhanced representativeness and balance without introducing systematic bias, thereby supporting robust and generalizable AI-driven FSC modelling.

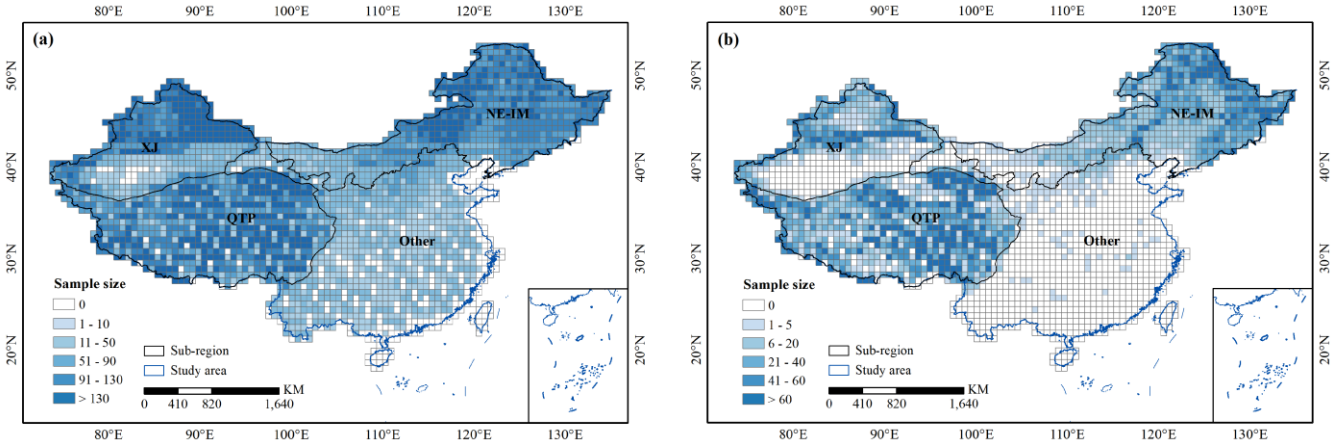

Figure 5: Spatial and snow condition completeness of the MODIS FSC sample dataset. (a) and (b) show the spatial distribution of the original and the post-QC AI-ready FSC samples, respectively.

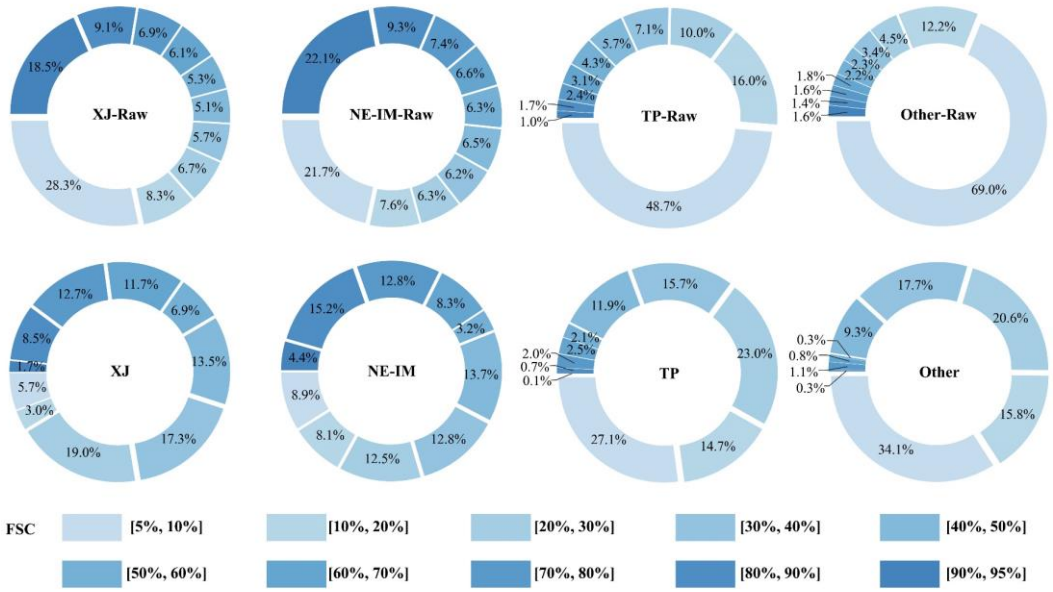

Figure 6: Distributions of tile-level mean FSC values before and after quality control. The upper and lower panels show the mean FSC distributions of the original and post-QC AI-ready samples, respectively, for the four sub-regions.

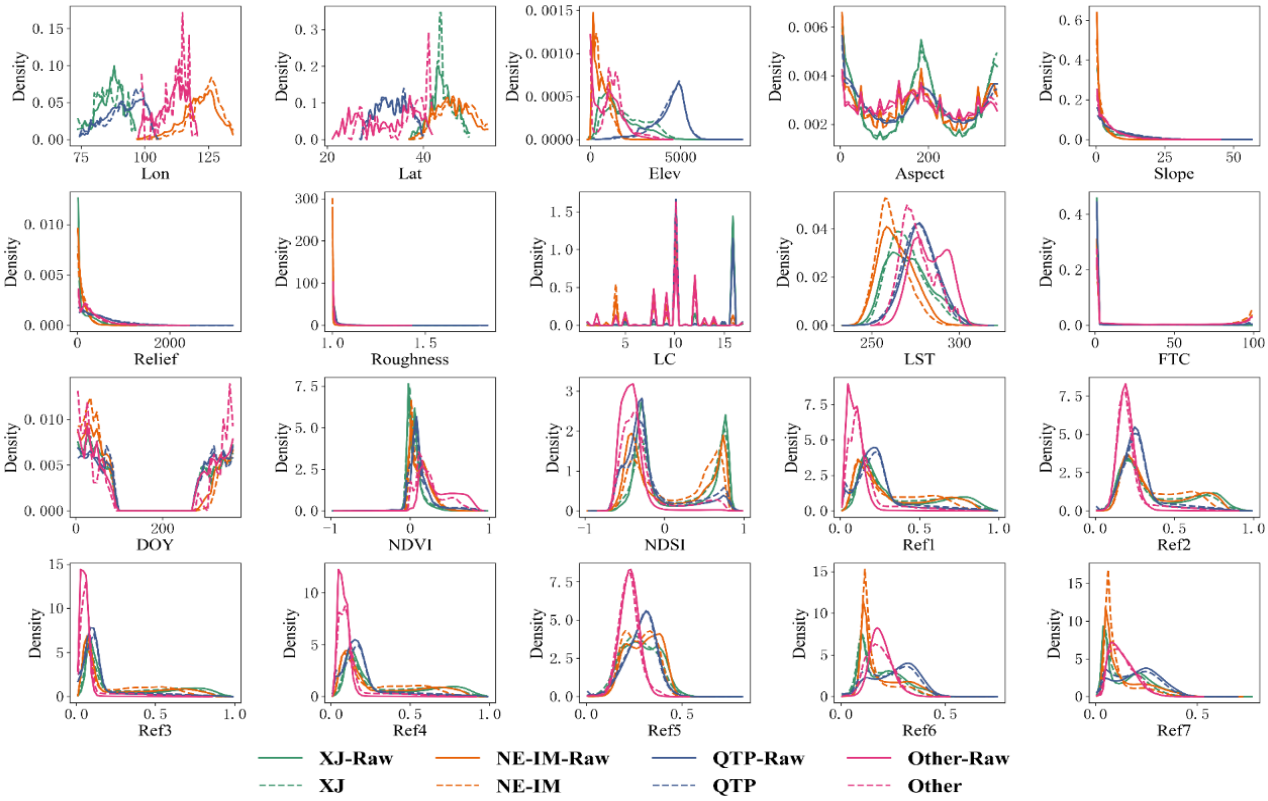

Figure 7: Kernel density distributions of the original and post-QC MODIS FSC samples

3.3.2 Physical Consistency

The quality of the AI-Ready FSC dataset was further assessed in terms of the physical consistency between reference FSC and its associated predictor variables. At the tile level, we quantified the pairwise correlations between FSC and all feature variables, as well as the inter-feature relationships (Fig. 8). The correlation structure exhibits clear and physically interpretable patterns. FSC shows strong positive correlations with NDSI and visible reflectance bands sensitive to snow, and negative correlations with LST and shortwave-infrared reflectance, consistent with the spectral and thermal properties of snow-covered surfaces. Elevation is positively associated with FSC, while slope- and roughness-related variables exhibit moderate but coherent relationships, reflecting topographic modulation of snow accumulation and persistence. Meanwhile, the strong correlations among spectral bands and between NDSI and reflectance confirm the internal radiometric consistency of the dataset. These statistically coherent relationships indicate that the feature-FSC matchups are not arbitrary associations but conform to well-established snow-environment interactions, providing a physically meaningful learning space for AI-based FSC modelling. Figure 9 further illustrates a representative sample, showing the spatial patterns of FSC together with key driving variables. The spatial correspondence between high FSC and high elevation, low LST, and high NDSI is clearly visible, whereas low FSC coincides with warmer temperatures, lower elevations, and reduced snow-sensitive spectral responses. This spatial

coherence complements the statistical evidence and demonstrates that the dataset preserves physical consistency not only in statistical, but also at the level of individual samples.

Such physical consistency is essential for robust AI learning, as it constrains models to capture meaningful snow-environment relationships. These results confirm that ChinaAI-FSC provides reliable and physically consistent feature-FSC matchups for ML and DL applications.

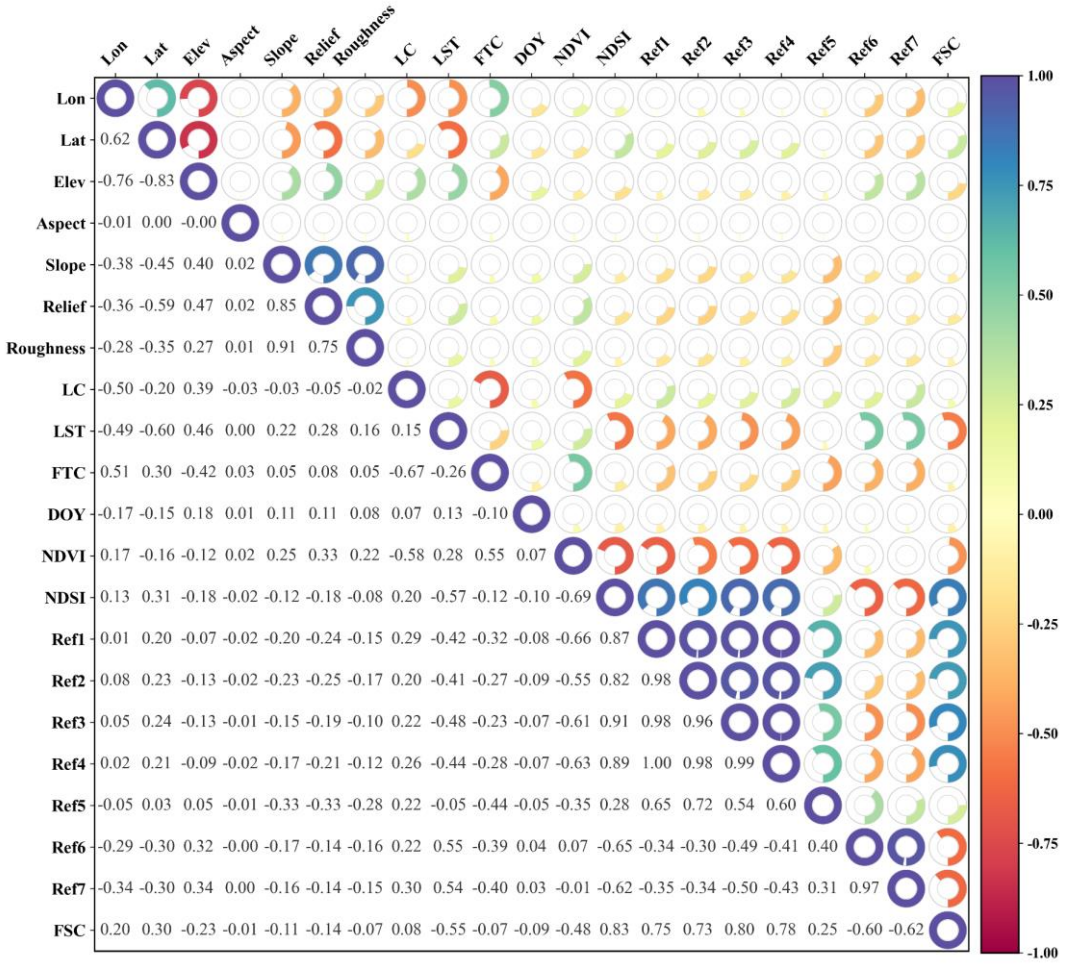

Figure 8: Correlations between FSC and all feature variables

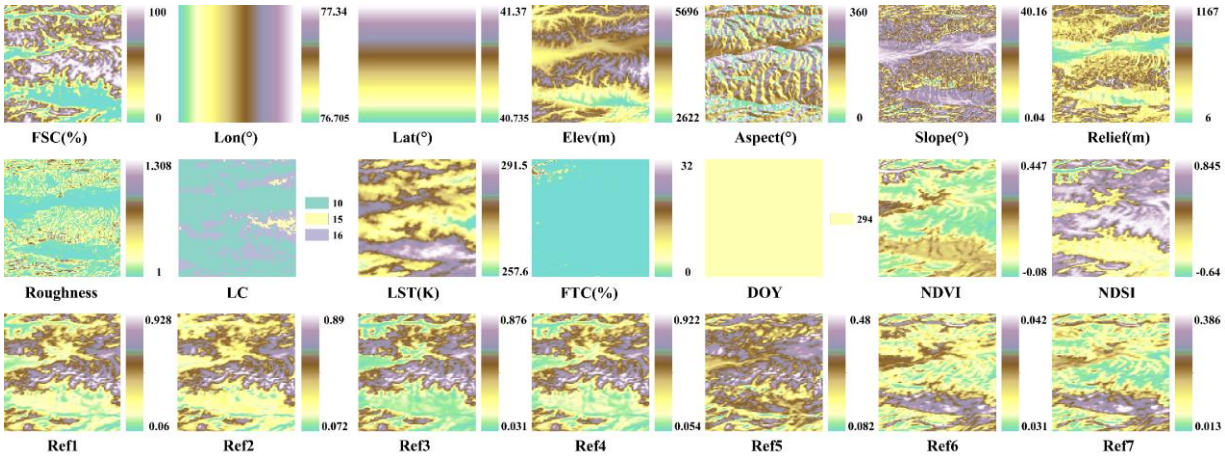

Figure 9: The spatial patterns of FSC and corresponding 20 input features for the example sample "201611204_r25c40_4", as illustrated in Figure 3.

3.3.3 Independent Validation Using In-situ Snow Depth

To further assess the physical reliability of the reference FSC used in ChinaAI-FSC, we performed an independent validation using in-situ SD observations. Station SD records were spatially matched to MODIS-scale pixels and temporally aligned with the corresponding satellite acquisition dates. Snow presence was defined as SD > 1 cm and FSC ⩾ 15%, following commonly adopted criteria in snow validation studies (Painter et al., 2009; Zhang et al., 2019). In total, 5016 independent SD-FSC validation pairs were obtained.

Based on confusion-matrix analysis (Table 6), the reference FSC exhibits strong agreement with in-situ observations over the entire study area, with an overall accuracy (OA) of 0.944. To further examine performance under different surface conditions, the validation samples were stratified according to land-cover type and terrain complexity into three categories: forested, mountainous (elevation ⩾ 2500 m or local relief > 200 m), and general (remaining non-forested regions). High consistency is maintained across all categories. The highest agreement is observed in mountainous regions (OA = 0.970), indicating that the high-resolution reference construction effectively captures terrain-modulated and heterogeneous snow patterns. However, it should be noted that the number of validation samples in mountainous areas is relatively limited, which may introduce additional uncertainty and partially inflate the estimated accuracy. In forested regions, the agreement remains high (OA = 0.906), although slightly lower than in non-forested areas. This reduction is consistent with canopy occlusion effects that weaken the optical snow signal in forest environments (Hall and Riggs, 2007; Metsämäki et al., 2012). Similarly, the smaller sample size in forested regions may contribute to a variability in the reference FSC accuracy.

Overall, this independent validation confirms that the reference FSC used in ChinaAI-FSC is physically reliable across diverse land-cover and terrain conditions, while also highlighting that accuracy in complex environments should be interpreted in the context of available in-situ sample density.

Table 6. Confusion-matrix-based independent validation of reference FSC using in-situ snow depth observations (2013-2020)

	TP	FP	FN	TN	Total Number	OA
General	2332	53	158	999	3542	0.940
Mountainous	41	15	18	1018	1092	0.970
Forested	206	17	19	140	382	0.906
China	2579	85	195	2157	5016	0.944

Note: TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative, respectively. OA represents overall accuracy.

3.3.4 AI-Readiness Evaluation Protocol

To rapidly assess the AI-readiness of Earth observation datasets, the U.S. National Oceanic and Atmospheric Administration (NOAA) proposed a four-tier maturity model (Levels 0-3), corresponding to *Not AI-Ready*, *Minimal*, *Intermediate*, and *Optimal* readiness, respectively (Christensen, 2020). This model assesses datasets from three perspectives, i.e., *data consistency, data accessibility, and metadata*, clearly defines the characteristics of each maturity level, providing a concise yet effective framework for AI-readiness assessment. Evidently, our AI-Ready MODIS FSC dataset demonstrably achieves the *highest maturity level (Level 3)* across all three dimensions. However, NOAA's model represents a generalized evaluation scheme and does not fully capture the multi-dimensional characteristics of AI-ready geospatial training datasets, particularly those involving multi-source harmonization, hierarchical spatial organization, and feature-target usability for AI, which are critical for FSC modelling and many other Earth system variables.

To address these limitations, we refined and extended NOAA's framework by introducing the "Four Layers-Four Domains-Fifteen Attributes" (4L-4D-15A) Evaluation Protocol (Table 7). This framework enables a granular, multi-perspective assessment of AI-readiness for geospatial training datasets, systematically characterizing them across four complementary dimensions: Data, Information, System, and Application. The fifteen attributes provide a structured basis for evaluating key properties such as data consistency, traceability, spatial organization, and algorithmic usability, which are essential for AI-based FSC modelling.

Using this FSC-oriented AI-readiness assessment, ChinaAI-FSC is found to achieve a high level of readiness across all four layers, indicating that the dataset provides internally consistent, well-documented, and spatially organized feature-FSC matchups suitable for training and evaluating machine learning and deep learning models. This assessment is intended to characterize the quality and usability of the FSC samples themselves, rather than to describe data services, platforms, or infrastructure.

Although demonstrated here for FSC, the 4L-4D-15A protocol is formulated around fundamental properties of training-sample-based Earth observation datasets, including multi-source harmonization, hierarchical spatial structure, and feature-target consistency. It is therefore applicable to other geophysical variables (e.g., soil moisture, vegetation, land surface temperature) that rely on similar multi-source training data, providing a methodological foundation for evaluating AI-readiness and scientific reliability across Earth observation applications.

Table 7: AI-Ready data characterization evaluation protocol of "Four Layers-Four Domains-Fifteen Attributes" (4L-4D-15A).

Layer	Domain	Attribute	Core Description
I. Data	Data Engineering	1. Preparation & Cleaning	Pre-cleaned, standardized, and ready for direct AI workflows.
		2. Multi-source Integration	Integrated from multi-source and multi-modal datasets with unified spatiotemporal alignment.
		3. Structure & Format	Structured and standardized formats for efficient AI processing.
II. Information	Information Integrity	4. Metadata & Annotation	Comprehensive metadata and semantically consistent annotations.
		5. Quality & Integrity	Ensures spatial-temporal-physical consistency with rigorous QC.
		6. Provenance & Traceability	Full provenance records for transparency and reproducibility.
		7. Timeliness	Frequently updated and temporally consistent.
III. System	System Interoperability	8. Accessibility	Accessible via APIs or open data services.
		9. Interoperability	Compliant with FAIR and OGC standards for system interoperability.
		10. Scalability	Designed for scalable and distributed AI computation.
		11. Reusability	Accompanied by clear documentation and reuse licensing.
IV. Application	AI Adaptability & Ethics	12. AI-task Adaptability	Optimized for AI tasks (classification, regression, segmentation) with balanced samples.
		13. Computational Efficiency	Optimized for HPC and GPU-based AI processing.
		14.Privacy & Ethics Compliance	Compliant with data privacy and ethical standards.
		15. Sustainability & Maintenance	Version-controlled and maintained for long-term sustainability.

3.4 Dataset Organization and Partitioning

To support robust evaluation of AI-based FSC models, the ChinaAI-FSC dataset is partitioned according to spatial and temporal considerations that explicitly target model generalization. The partitioning strategy is designed to minimize information leakage while preserving the intrinsic heterogeneity of snow processes across China.

Spatially, the dataset is divided into four subregional domains, XJ, NEIM, TP, and Other, corresponding to the major snow-climate regimes of mainland China. Within each subregion, samples are partitioned into training, validation, and testing subsets following a spatially disjoint 2:1:1 spatial ratio (Fig. 10). This spatial separation reduces the impact of spatial autocorrelation, ensuring that model evaluation reflects the ability to generalize across distinct geographic areas rather than exploiting local similarity. Temporally, each spatial subset spans 22 snow seasons. Importantly, samples that are spatially adjacent but belong to different snow seasons are treated as independent realizations of snow-environment interactions, reflecting the strong interannual variability of snow cover. By combining spatial independence with long-term temporal coverage, the adopted partitioning strategy enables rigorous assessment of both spatial and interannual generalization. It provides a reliable basis for benchmarking AI models under diverse snow conditions, while avoiding overly optimistic performance estimates caused by spatial or temporal leakage.

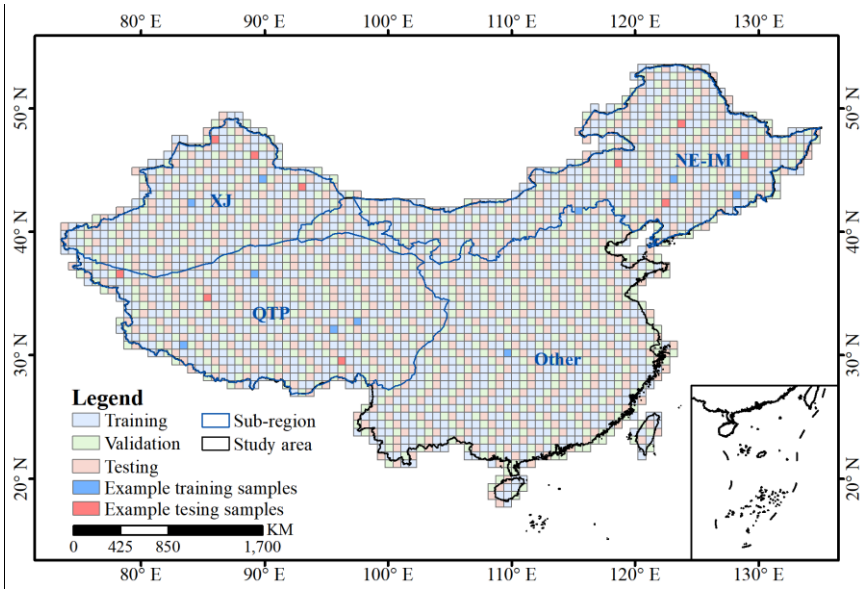

Figure 10: Spatial distribution of samples assigned to training, calibration, and testing dataset with a 2:1:1 ratio across the study area. The indicated "Example training samples" and "Example testing samples" correspond to the representative samples analysed in Section 4.1 and shown in Figure 11 and 12.

4. Demonstration Applications Using the AI-Ready FSC Dataset

To demonstrate the quality, reliability, and applicability of the AI-Ready FSC dataset, three representative applications were conducted: (1) benchmark modelling using relatively simple, well-established algorithms without extensive hyperparameter tuning to evaluate dataset robustness and usability, (2) assessment of MODIS FSC product accuracy using the dataset, and (3) large-scale seamless FSC mapping over China to examine the dataset's representativeness across the study region and the generalization ability of AI models. These experiments aimed to provide a transparent and robust baseline rather than optimized prediction performance.

4.1 Benchmark Modelling

To evaluate the quality, reliability, and applicability of the AI-Ready FSC dataset, a set of benchmark models was established using samples from the 2021-2022 snow season. Six representative algorithms, i.e., ANN, SVR, RF, CNN, UNet, and ResNet, were implemented to assess the dataset's performance across different modelling paradigms. Each model was trained, validated, and tested using spatially independent data splits derived from the AI-Ready sample structure, thereby ensuring that the performance evaluation reflects the dataset's capability for geographic generalization rather than random sample fitting. Model performance was assessed using three standard metrics: root mean square error (RMSE), mean absolute error, and coefficient of association (R). The detailed architectural configurations, hyperparameter settings, and training strategies for all benchmark models are summarized in Table 8, providing a reproducible foundation for model replication and comparative analysis. It is

important to emphasize that all benchmark models were implemented using their canonical architecture and commonly adopted hyperparameter settings, without any deliberate optimization of network structure, parameter tuning, or learning algorithm modification. This design choice ensures that the benchmarking process isolates and highlights the intrinsic quality, consistency, and representativeness of the AI-Ready FSC samples, rather than reflecting model-specific tuning effects. Accordingly, the resulting performance metrics offer an objective and unbiased evaluation of the dataset's robustness, reliability, and applicability for diverse AI-driven snow-cover modelling frameworks. These results thus serve as a baseline reference for subsequent algorithmic development, dataset intercomparison, and large-scale AI-readiness benchmarking within the cryosphere research community.

Table 8: The architectural configurations and parameter settings of the six benchmark models

Method	Model Structure	Model Parameters
ANN	Three hidden layers with 128, 64, and 32 neurons	Optimizer: Adam; Learning rate: 1e-3; Loss: MSE; Max epochs: 200; Early stopping patience: 20
SVR	Nonlinear mapping with radial basis function (RBF) kernel	Penalty (C): 10; Kernel parameter: scale
RF	Ensemble of decision trees trained on bootstrap strategy, outputs are aggregated by averaging	Total trees: 200; Max features: sqrt(200); Max depth: 20;
CNN	3× (Conv3×3, BatchNorm2d, ReLU)	Optimizer: Adam; Batch size: 32; Learning rate: 1e-3; Loss: MSE; Max epochs: 200; Early stopping patience: 20
UNet	Three encoder layers, bottleneck, 3 decoder layers with skip connections; each conv block: 2× (Conv3×3, BatchNorm2d, ReLU)	Optimizer: Adam; Batch size: 32; Learning rate: 1e-3; Loss: MSE; Max epochs: 200; Early stopping patience: 20
ResNet	Four residual blocks, each combining convolutional layers (Conv3×3, BatchNorm2d, ReLU) with skip connections each residual block	Optimizer: Adam; Batch size: 32; Learning rate: 1e-3; Loss: MSE; Max epochs: 200; Early stopping patience: 20

Figures 11 and 12 illustrate the FSC distributions estimated by six benchmark models for ten representative example samples from distinct spatial locations (Fig. 10), selected from the training and testing datasets, respectively, spanning the entire FSC range from 0 to 100% in 10% intervals. Overall, all models successfully capture the primary spatial patterns of snow cover, demonstrating the capability of both ML and DL approaches to reproduce tile-scale FSC variability. Among the point-based ML models, ANN, SVR, and RF effectively approximate spatial heterogeneity in FSC, although SVR tends to exhibit slightly weaker performance, particularly in complex terrain or intermediate FSC ranges. Estimation errors tend to increase in snow-rich and structurally complex regions, mainly due to spectral saturation over highly reflective snow surfaces, canopy occlusion that limits the visibility of underlying snow, mixed-pixel effects arising from subpixel snow-land heterogeneity, terrain-induced illumination variability, spatial variations in forest structure, and snow grain size-dependent changes in spectral response. In contrast, the tile-based DL models, i.e., CNN, UNet, and ResNet, demonstrate enhanced capability in capturing coherent spatial structures and snow boundaries, yielding FSC distributions more consistent with reference values and effectively reducing local discrepancies, especially in continuous snow-covered regions. Comparative analysis between training and testing results reveals slightly higher accuracy for training samples, reflecting their stronger spatial representativeness; however, DL models maintain robust generalization and transferability across independent test samples. Collectively, these findings confirm that both point- and tile-scale AI models can effectively reproduce FSC spatial patterns, with DL architectures offering superior accuracy and spatial coherence. More importantly, the results highlight that the

constructed AI-ready FSC dataset provides a physically consistent, statistically representative, and algorithmically versatile foundation for training and evaluating AI-based snow-cover models, supporting large-scale, high-precision snow mapping across heterogeneous regions such as China.

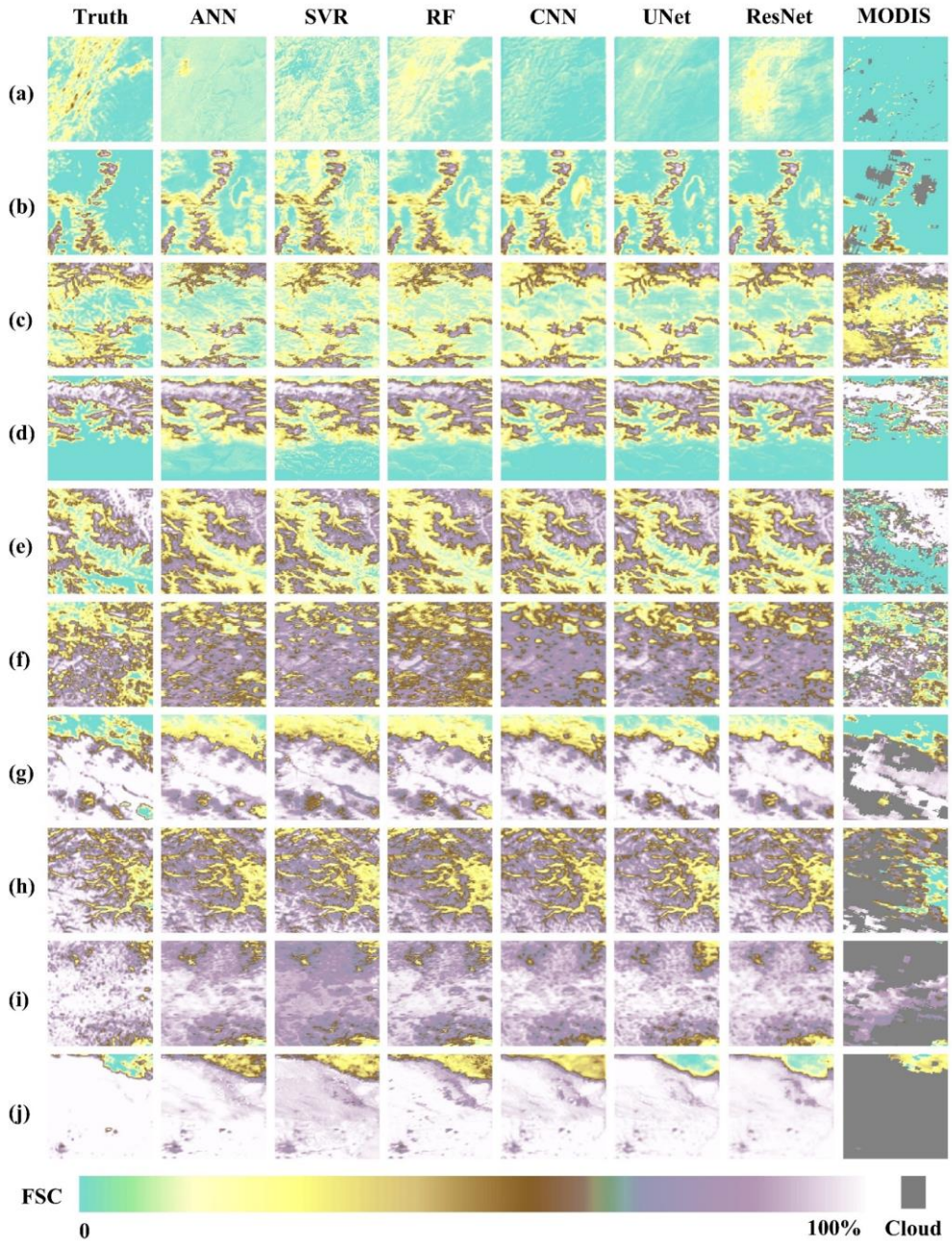

Figure 11: FSC distributions of the reference truth (sample labels, first column), six benchmark models, and the MODIS standard FSC product (last column) for example training samples. Panels (a)-(j) represent different snow cover conditions, spanning the full FSC range of [0-10%], [10-20%], …, [90-100%], respectively.

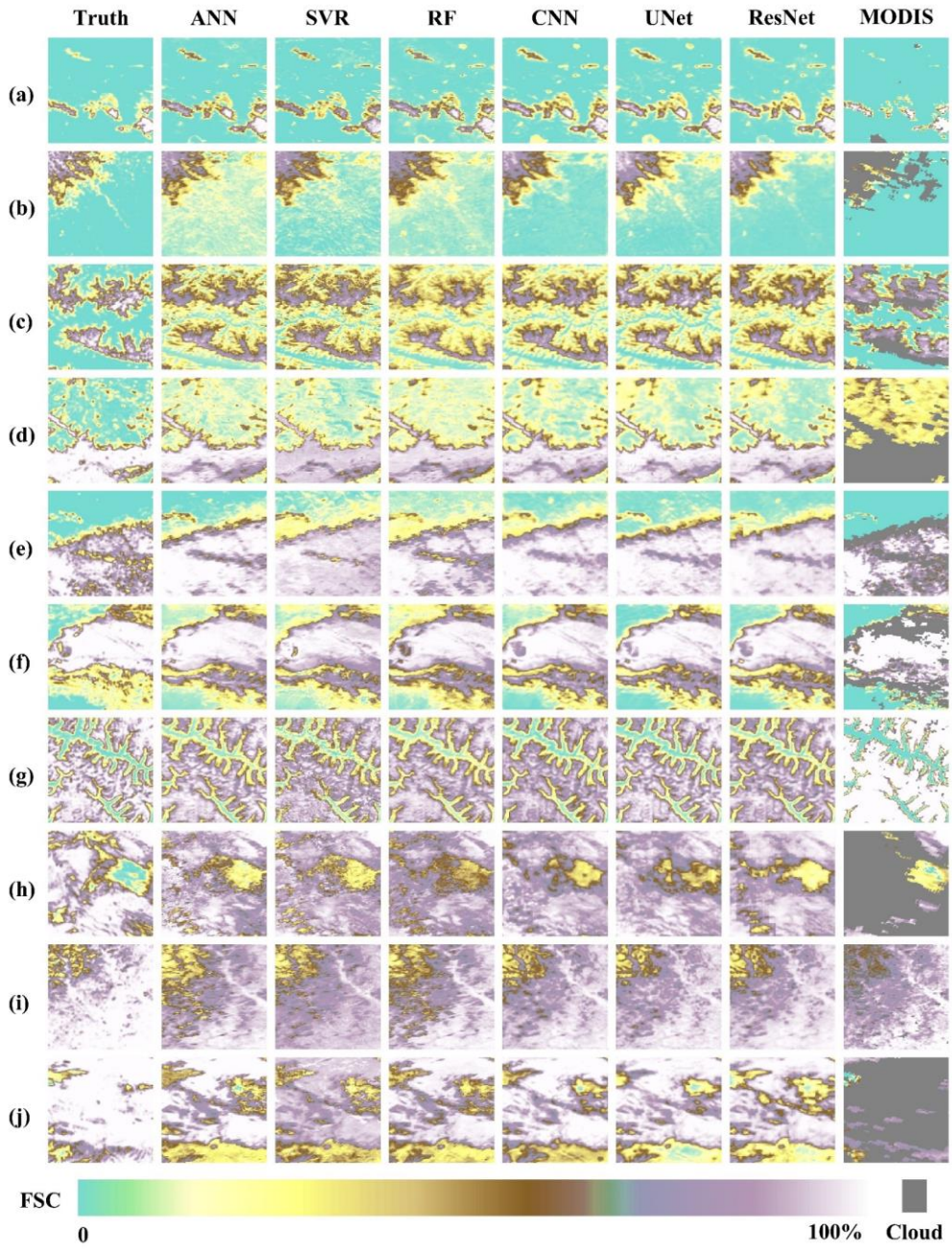

Figure 12: Same as Figure 9, but samples from the testing dataset.

Figure 13 presents the performance comparison of the six benchmark models on both the training and testing datasets. On the training set, overall performance was robust, with the ResNet model achieving the best results (R = 0.89, RMSE = 11.69%, MAE = 8.13%), followed closely by UNet. On the testing set, the UNet model achieved the highest performance (R = 0.86, RMSE = 14.21%, MAE = 9.57%), slightly outperforming ResNet (R = 0.86, RMSE = 14.33%, MAE = 9.84%). The point-

scale ML models (ANN, SVR, and RF) generally performed worse than the tile-scale DL models (CNN, UNet, and ResNet), although all achieved satisfactory accuracy levels. Overall, the UNet model demonstrated the best balance between predictive accuracy and stability. A more detailed quantitative evaluation across different FSC intervals (Table 9) further supports this conclusion. Under low (FSC < 20%) or high (FSC > 70%) snow-cover conditions, the surface environment is relatively homogeneous, dominated either by bare-ground reflectance or by continuous snow-covered surfaces. This spectral uniformity minimizes within-window variability, resulting in more stable feature-response relationships and consequently higher model accuracy. In contrast, at intermediate FSC levels (20%-70%), the coexistence of snow, vegetation, and exposed soil increases sub-pixel heterogeneity and introduces nonlinear interactions among spectral, thermal, and topographic factors. These conditions amplify model uncertainty and lead to a noticeable decrease in prediction accuracy for all models. In short, the results indicate that model performance is governed more by the quality, representativeness, and internal consistency of the training data than by the specific model architecture. When the training samples adequately capture the full range of spectral, topographic, and environmental variability, even relatively simple models can achieve high predictive accuracy. This finding underscores that the intrinsic data quality and representativeness of the AI-ready FSC samples are the primary determinants of model generalization and robustness.

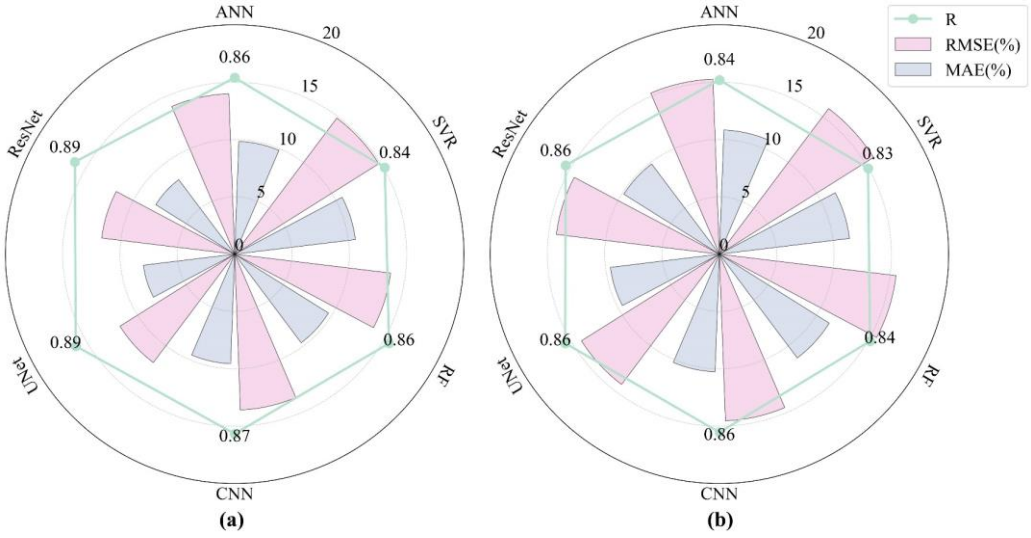

Figure 13: Performance comparison of six benchmark models on the training and testing FSC sample datasets. Panels (a) and (b) show the evaluation results for the training and testing subsets, respectively.

4.2 MODIS Standard FSC Product Accuracy Assessment

Furthermore, all reference FSC values from the 2021-2022 snow season were employed to independently validate the accuracy of the MODIS standard FSC product, in which FSC estimates are derived from the NDSI values in the MOD10A1 product by directly applying the official fitting coefficients from Salomonson and Appel (2004). As summarized in Table 9, the quantitative comparison shows that the MODIS standard FSC product achieves reasonable accuracy, with performance comparable to the point-scale ML models (ANN, SVR, and RF). However, it remains clearly inferior to the tile-scale DL

models (CNN, UNet, and ResNet), highlighting that spatially explicit architectures can more effectively capture contextual information and spatial continuity in FSC estimation. Moreover, since the MODIS standard algorithm retrieves FSC only under clear-sky conditions, the average effective coverage is typically below 50%, indicating that more than half of the surface area is cloud-obscured and thus excluded from estimation. This limitation substantially restricts its spatial completeness. In contrast, the AI-Ready FSC sample dataset constructed in this study exhibits robust performance under diverse atmospheric and surface conditions, ensuring high spatial and temporal continuity. The FSC distribution maps for representative sample tiles (Fig. 11 and 12) further support these findings, visually confirming the superior consistency and realism of the AI-based FSC estimates. Importantly, the large-scale validation involving 3686 independent samples, each corresponding to one Landsat or Sentinel-2 scene, ensures the statistical robustness and credibility of the evaluation, providing strong evidence for the high quality, reliability, and representativeness of the constructed dataset.

Table 9. Quantitative evaluation of six benchmark models and MODIS standard FSC product based on different FSC intervals

	FSC intervals	ANN	SVR	RF	CNN	UNet	ResNet	MODIS
RMSE (%)	[0,10]	9.93	10.20	9.55	9.24	8.44	8.47	8.99
	[10,20]	12.45	12.78	12.04	11.81	11.13	11.14	14.01
	[20,30]	15.00	15.70	15.00	14.59	13.59	13.64	15.35
	[30,40]	16.51	17.41	16.66	15.98	14.79	14.76	15.91
	[40,50]	17.56	18.55	17.78	17.06	15.67	15.70	16.51
	[50,60]	15.99	16.63	16.13	15.82	14.89	14.80	16.85
	[60,70]	14.59	15.27	14.56	13.65	12.95	12.90	15.66
	[70,80]	14.97	15.69	14.93	13.93	13.34	13.30	14.12
	[80,90]	15.01	15.98	14.97	14.21	12.61	12.61	13.92
	[90,100]	14.30	15.59	14.71	14.00	11.05	11.41	13.76
	Average	14.63	15.38	14.63	14.03	12.85	12.87	14.51
MAE (%)	[0,10]	6.72	7.04	6.50	5.91	4.82	5.21	6.04
	[10,20]	8.49	8.82	8.19	7.90	6.89	7.19	9.64
	[20,30]	10.82	11.30	10.92	10.39	9.29	9.49	10.68
	[30,40]	12.06	12.71	12.33	11.66	10.44	10.54	11.37
	[40,50]	12.99	13.73	13.32	12.61	11.25	11.38	12.10
	[50,60]	11.48	12.10	11.78	11.41	10.52	10.60	11.96
	[60,70]	10.12	10.91	10.33	9.63	8.95	9.03	11.20
	[70,80]	9.99	10.97	10.29	9.55	8.91	9.01	9.95
	[80,90]	9.91	11.23	10.27	9.67	8.33	8.51	10.45
	[90,100]	9.36	11.09	10.07	9.65	7.18	7.69	10.98
	Average	10.19	10.99	10.40	9.84	8.66	8.87	10.44
R	[0,10]	0.82	0.82	0.84	0.85	0.86	0.86	0.83
	[10,20]	0.84	0.83	0.85	0.85	0.87	0.87	0.80
	[20,30]	0.85	0.83	0.85	0.85	0.88	0.87	0.86
	[30,40]	0.85	0.83	0.85	0.86	0.88	0.88	0.89
	[40,50]	0.85	0.84	0.85	0.86	0.88	0.88	0.89
	[50,60]	0.88	0.87	0.88	0.89	0.90	0.90	0.90
	[60,70]	0.89	0.88	0.89	0.90	0.91	0.91	0.89
	[70,80]	0.84	0.83	0.84	0.86	0.87	0.87	0.87
	[80,90]	0.76	0.74	0.76	0.77	0.81	0.81	0.76
	[90,100]	0.64	0.63	0.66	0.67	0.75	0.74	0.66
	Average	0.91	0.90	0.91	0.92	0.93	0.93	0.92

4.3 Large-Scale FSC Mapping across China

Figure 14 illustrates the spatial FSC distribution across China on three representative dates, i.e., October 1, 2021 (accumulation period), January 18, 2022 (stable period), and March 8, 2022 (ablation period), as estimated by the Unet model developed in Section 4.1. The retrieved patterns exhibit clear spatial continuity and physically consistent gradients with respect to both latitude and elevation. During the early accumulation phase (October 1, 2021), snow cover appears primarily over high-altitude regions, including the QTP, the Tianshan and Altai Mountains, and the northeastern highlands (Greater Khingan Range and Changbai Mountains), while most lowlands and southeastern coastal regions remain snow-free. By January 18, 2022, the snow extent reaches its annual maximum, forming an almost continuous snow belt across the QTP and northern China, with FSC values exceeding 0.8 in cold and high-elevation zones. On March 8, 2022, a clear retreat of snow cover is observed, characterized by rapid melting in low- and mid-latitude areas, whereas residual snow persists in high mountains and northern forests, depicting a physically consistent seasonal evolution.

Spatially, the Unet-derived FSC maps exhibit smooth transitions and coherent snowline boundaries, indicating that the model effectively captures contextual terrain information and suppresses pixel-level noise. The altitude-FSC relationship remains nearly monotonic, with FSC increasing systematically with elevation, further validating the physical realism of the estimates. Minor discontinuities and isolated patches occur mainly along steep slopes or shaded terrains, where topographic shadows or mixed-pixel effects may distort spectral responses. When compared with MODIS clear-sky FSC observations, the Unet results display strong spatial agreement across most regions. Based on two randomly selected representative tiles for each date (Table 10), the UNet-derived FSC achieves mean R above 0.92, with an average RMSE of 10.75% and MAE of 6.46%, notably outperforming the corresponding MODIS product (RMSE=4.41%, MAE=9.34%, R= 0.91). Although slight biases persist, such as minor underestimation in high-FSC areas and overestimation in low-FSC zones, the UNet model demonstrates clear advantages in spatial continuity, snowline delineation, and temporal consistency. Overall, these results confirm that the Unet-based FSC mapping provides physically reliable, spatially coherent, and temporally consistent characterization of snow-cover dynamics across China. More importantly, the constructed AI-Ready FSC sample dataset establishes a robust, standardized, and scalable foundation for training, validating, and benchmarking advanced AI models, thereby enabling high-precision, large-scale snow-cover mapping over complex and heterogeneous regions such as China.

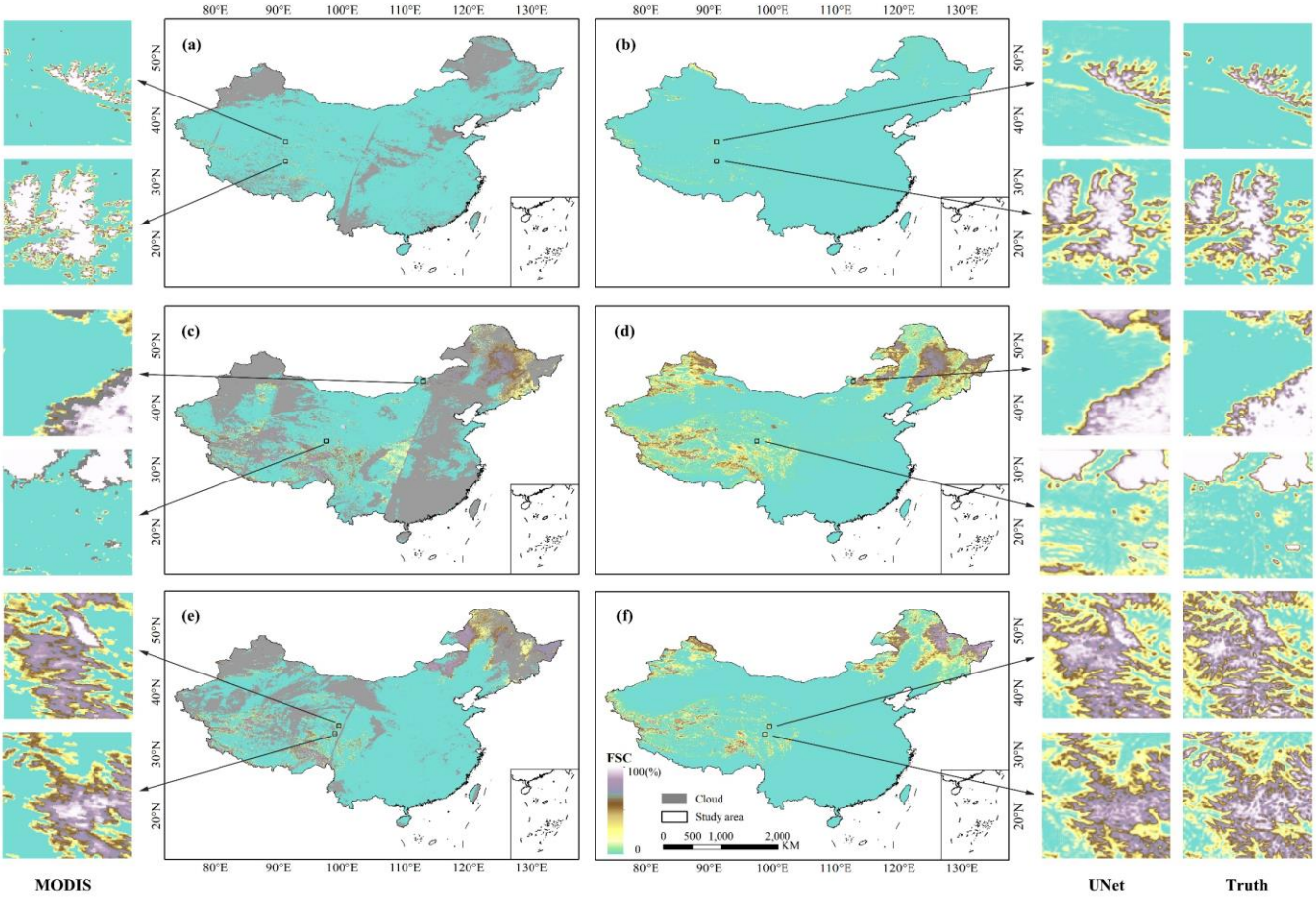

Figure 14: Spatial distribution of FSC across China on three representative dates derived from the MODIS FSC product (subplots (a), (c), and (e)) and the UNet model estimates (subplots (b), (d), and (f)). Panels (a)-(b) correspond to October 1, 2021 (accumulation period), (c)-(d) to January 18, 2022 (stable period), and (e)-(f) to March 8, 2022 (ablation period). Insets show example tiles, i.e., r30c28, r25c28, r42c62, r27c38, r27c41, and r25c40 (from top to bottom), highlighting local-scale improvements in snow delineation and spatial coherence.

Table 10. Quantitative evaluation of randomly selected example samples on three representative dates, i.e., October 1, 2021 (accumulation period), January 18, 2022 (stable period), and March 8, 2022 (ablation period), from UNet model and standard MODIS FSC product.

Model	Sample (tile)	Reference FSC (%)	RMSE (%)	MAE (%)	R
UNet	r30c28	6.31	5.89	2.74	0.97
	r25c28	32.08	8.01	5.18	0.98
	r42c62	26.22	16.38	7.82	0.94
	r27c38	22.77	10.94	6.38	0.97
	r27c41	49.38	10.44	7.81	0.94
	r25c40	44.81	12.85	8.85	0.93
MODIS FSC	r30c28	6.31	7.63	3.42	0.96
	r25c28	32.08	12.81	7.33	0.96
	r42c62	26.22	17.24	9.34	0.92
	r27c38	22.77	12.54	8.03	0.95
	r27c41	49.38	18.28	14.28	0.90
	r25c40	44.81	17.98	13.64	0.91

5. Discussion

5.1 Methodological implications of AI-ready FSC dataset construction

This study treats the construction of an AI-ready FSC dataset not merely as a data compilation task, but as a methodological problem centered on learning validity. Rather than optimizing for a specific retrieval algorithm or application, ChinaAI-FSC is designed to ensure that feature-FSC matchups are physically consistent, structurally coherent, and statistically suitable for data-driven modelling. From this perspective, dataset construction itself becomes a critical determinant of model robustness, interpretability, and generalization.

A key methodological implication is the shift from accuracy-oriented quality control toward learning-oriented quality assessment. By enforcing consistency between FSC references and their driving variables (e.g., terrain, temperature, and spectral indicators), and by evaluating samples at both pixel and tile levels, the dataset constrains AI models to learn within physically plausible regimes. This reduces the risk of spurious correlations that may arise from purely statistical filtering and enhances robustness under heterogeneous snow conditions.

Furthermore, the standardized feature-FSC pairing and hierarchical spatial organization highlight the importance of structural coherence in AI-ready geospatial datasets. Instead of tailoring the dataset to a particular model architecture, ChinaAI-FSC is constructed to remain model-agnostic, supporting a wide range of learning paradigms including point-based regression, spatial modelling, and uncertainty-aware approaches. These considerations emphasize that AI-readiness is fundamentally a methodological attribute of the dataset rather than a byproduct of downstream model selection.

5.2 Data availability constraints and sample imbalance

The construction of ChinaAI-FSC is fundamentally constrained by the availability of high-quality, near-cloud-free Landsat and Sentinel-2 observations, which remain uneven in both space and time, particularly over mountainous, high-latitude, and persistently cloudy regions. Under these observational constraints, the dataset was assembled by exhaustively collecting all reference-quality observations that satisfied strict quality criteria, rather than by imposing an explicit sampling or stratification scheme aimed at balancing snow conditions.

As a result, although the adopted quality control procedures substantially enhance physical consistency and learning validity, a certain degree of sample redundancy and imbalance remains, most notably across different FSC intervals. Extremely sparse and extremely dense snow conditions are relatively underrepresented compared to moderate FSC ranges. This imbalance reflects the intrinsic spatiotemporal characteristics of snow cover and the inherent limitations of optical remote sensing, rather than deficiencies in the dataset construction methodology.

Importantly, ChinaAI-FSC does not assume balanced sample distributions as a prerequisite for AI-based FSC modelling. Instead, it provides a transparent and well-characterized reference sample repository that explicitly exposes real-world sample imbalance under current observational constraints. In this sense, the dataset serves not only as training material but also as a

foundation for investigating how sample imbalance and redundancy influence AI-based FSC estimation, and for developing mitigation strategies such as adaptive sampling, sample reweighting, and physically consistent data augmentation.

5.3 Uncertainty and limitations in FSC modelling over complex surfaces

Uncertainty in FSC modelling over complex surfaces primarily arises from subpixel snow heterogeneity, terrain-induced illumination effects, and forest canopy interactions, which jointly affect both FSC reference estimation and predictor variables. Subpixel snow is widespread during accumulation and melt periods and in transitional climate zones, where mixed snow-land signals dominate optical observations and introduce inherent ambiguity in FSC retrieval (Salomonson and Appel, 2004; Painter et al., 2009; Rittger et al., 2013).

In mountainous regions, complex terrain further amplifies uncertainty through variations in slope, aspect, and cast shadows, which alter surface reflectance and thermal conditions independently of snow presence (Klein and Barnett, 2003). In forested environments, canopy occlusion reduces the visibility of underlying snow and weakens spectral snow signals, leading to systematic underestimation of FSC in optical products (Hall and Riggs, 2007; Metsämäki et al., 2012). These challenges are widely recognized in snow remote sensing and are not specific to the ChinaAI-FSC dataset.

By using high-resolution Landsat and Sentinel-2 imagery as reference, ChinaAI-FSC captures patchy and heterogeneous snow patterns that are often poorly represented in binary snow products, thereby supporting fractional rather than categorical snow modelling. At the same time, residual uncertainties associated with terrain shadowing, canopy obscuration, and cross-sensor differences remain unavoidable, particularly in complex environments (Dietz et al., 2012). While topographic and forest-related features are incorporated to characterize these conditions, such uncertainties propagate into both the FSC reference and the feature space.

Rather than eliminating these uncertainties, the dataset is designed to make them transparent and diagnosable. The explicit feature-target organization, standardized tiling strategy, and consistency-based quality control facilitate uncertainty-aware modelling approaches, including ensemble learning, probabilistic FSC estimation, and sensitivity analysis. In this context, robustness is achieved not through the absence of uncertainty, but through explicit characterization and compatibility with uncertainty-aware AI frameworks.

5.4 Generalizability of the AI-readiness evaluation framework

Beyond the dataset itself, a key contribution of this study is the proposed AI-readiness evaluation framework, which shifts attention from product-level accuracy to dataset-level learning validity. Existing AI-readiness concepts often emphasize data accessibility and metadata completeness, but provide limited guidance on whether a dataset is structurally suitable for data-driven learning.

The introduced multi-dimensional framework evaluates datasets from data, information, system, and application perspectives, explicitly incorporating criteria related to feature–target coherence, physical consistency, and usability in AI workflows. While ChinaAI-FSC serves as a concrete implementation, the framework is not specific to snow cover. Its principles are applicable

to other Earth observation variables, including vegetation properties, soil moisture, land surface temperature, and hydrological states, where AI methods increasingly rely on large, multi-source datasets.

By decoupling AI-readiness evaluation from any specific model or platform, this framework provides a transferable methodological reference for constructing and assessing AI-ready geophysical datasets under realistic observational constraints. In this sense, the framework complements the dataset by offering a standardized and scientifically grounded approach to evaluating the robustness, usability, and learning validity of Earth observation data products.

6 Code and data availability

The ChinaAI-FSC dataset (Hou et al., 2025) is publicly available at the National Tibetan Plateau Data Center (TPDC) at https://doi.org/10.11888/Cryos.tpdc.303034 (also accessible via https://cstr.cn/18406.11.Cryos.tpdc.303034) and from Zenodo at https://doi.org/10.5281/zenodo.17707386, hosted on an open-access repository that supports persistent identifiers and version control. The full dataset, including feature variable tiles, reference FSC tiles, metadata files, and documentation, can be accessed and downloaded under a Creative Commons Attribution (CC BY 4.0) license.

All associated code, including data reading examples, processing scripts, and benchmark modeling workflows, is publicly available in our GitHub repository: https://github.com/houjin0503/AI-Ready-China-FSC. This ensures full transparency, traceability, and reproducibility of both the data generation and modelling processes.

7 Summary

This study presents ChinaAI-FSC, the first large-scale, AI-ready MODIS FSC sample dataset for mainland China, establishing a standardized, high-quality foundation for AI-based snow monitoring and modelling. The dataset integrates multi-source satellite observations with rigorous physical and quality control procedures, while implementing a novel "Four Layers-Four Domains-Fifteen Attributes" (4L-4D-15A) evaluation framework to ensure spatiotemporal representativeness, physical consistency, environmental completeness, and metadata standardization.

By fully embodying the AI-readiness paradigm, ChinaAI-FSC provides a reproducible and interoperable basis for the development, benchmarking, and intercomparison of ML and DL models for FSC estimation. It enables large-scale and temporally consistent FSC mapping, facilitates cross-sensor validation (e.g., between MODIS and VIIRS), and supports multi-scale AI workflows from local retrieval to continental-scale modelling. Beyond serving as a benchmark dataset, ChinaAI-FSC also contributes to methodological innovation in several key areas: (i) feature optimization and AI model interpretability, by providing richly annotated feature-response pairs suitable for sensitivity and explainability analyses; (ii) physics-AI hybrid modelling, by offering high-quality training and validation data that enable the integration of physical constraints into data-driven frameworks; and (iii) data assimilation and Earth system modelling, by supplying consistent observational inputs for model calibration and coupling.

In short, this work contributes a methodological innovation to the field of snow monitoring: the establishment of a continental-scale, AI-ready FSC dataset with standardized construction, multi-layer quality control, and a formal evaluation framework. This dataset paradigm emphasizes reproducibility, cross-regional generalization, and AI-readiness, and is explicitly designed to support machine learning and deep learning applications. The novelty lies in dataset methodology and standardization, rather than in proposing new retrieval algorithms.

While current limitations remain, such as imbalanced snow-condition samples, restricted spatial coverage over China, and residual uncertainties linked to cloud contamination, complex terrain, and sensor discrepancies, these also define clear pathways for future enhancement. Upcoming releases will focus on extending spatiotemporal coverage, enriching feature diversity, and strengthening cross-sensor harmonization to further improve dataset completeness and continuity. Overall, ChinaAI-FSC represents a versatile, open, and FAIR-compliant resource that advances AI-driven snow monitoring and model development, enhances algorithmic robustness and interpretability, and supports regional-to-global assessments of cryosphere dynamics under a rapidly changing climate.

Funding

This work was supported by National Natural Science Foundation of China (Grant No. 42130113, 42371398, 42471434, and 42361060), and the program of the Key Laboratory of Cryospheric Science and Frozen Soil Engineering, CAS (No. CSFSE-ZZ-2409).

Competing interests

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

Author contribution

JH and CH conceived the study and designed the overall methodology. JH, MZ, and YZ developed the model code and conducted the simulations. XH, JG, and PD performed the formal analyses. CH provided key resources and secured project funding. JH prepared the original draft of the manuscript. All authors contributed to the review and editing of the final manuscript.

References

Azizi, A. H., Akhtar, F., Kusche, J., Tischbein, B., Borgemeister, C., and Oluoch, W. A.: Machine learning-based estimation of fractional snow cover in the Hindukush Mountains using MODIS and Landsat data, J. Hydrol., 638, 131579, https://doi.org/10.1016/j.jhydrol.2024.131579, 2024.

Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential impacts of a warming climate on water availability in snow-dominated regions, Nature, 438, 303-309, https://doi.org/10.1038/nature04141, 2005.

Chander, G., Hewison, T. J., Fox, N., Wu, X., Xiong, X., and Blackwell, W. J.: Overview of intercalibration of satellite instruments, IEEE Trans. Geosci. Remote Sens., 51, 1056–1080, https://doi.org/10.1109/TGRS.2012.2228654, 2013.

Christensen, T.: What is AI-Ready Open Data?, presented 22 October 2020, NOAA NESDIS/STAR, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, available at: https://www.star.nesdis.noaa.gov/star/documents/meetings/2020AI/presentations/202010/20201022_Christensen.pdf (last access: 25 October 2025), 2020.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., Skakun, S. V., and Justice, C.: The Harmonized Landsat and Sentinel-2 surface reflectance data set, Remote Sens. Environ., 219, 145-161, https://doi.org/10.1016/j.rse.2018.09.002, 2018.

Crawford, C. J., Roy, D. P., Arab, S., Barnes, C., Vermote, E., Hulley, G., Gerace, A., Choate, M., Engebretson, C., Micijevic, E., and Schmidt, G.: The 50-year Landsat Collection 2 archive, Sci. Remote Sens., 8, 100103, https://doi.org/10.1016/j.srs.2023.100103, 2023.

Czyzowska-Wisniewski, E. H., van Leeuwen, W. J., Hirschboeck, K. K., Marsh, S. E., and Wisniewski, W. T.: Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network, Remote Sens. Environ., 156, 403-417, https://doi.org/10.1016/j.rse.2014.09.026, 2015.

Dobreva, I. D. and Klein, A. G.: Fractional snow cover mapping through artificial neural network analysis of MODIS surface reflectance, Remote Sens. Environ., 115, 3355-3366, https://doi.org/10.1016/j.rse.2011.07.018, 2011.

Dietz, A. J., Kuenzer, C., Gessner, U., and Dech, S.:Remote sensing of snow – a review of available methods, Int. J. Remote Sens., 33, 4094–4134, https://doi.org/10.1080/01431161.2011.640964, 2012.

Dozier, J.: Spectral signature of alpine snow cover from the Landsat Thematic Mapper, Remote Sens. Environ., 28, 9-22, https://doi.org/10.1016/0034-4257(89)90101-6, 1989.

Dozier, J., Painter, T. H., Rittger, K., and Frew, J. E.: Time-space continuity of daily maps of fractional snow cover and albedo from MODIS, Adv. Water Resour., 31, 1515-1526, https://doi.org/10.1016/j.advwatres.2008.08.011, 2008.

Essery, R., and Pomeroy, J. W.: Vegetation and topographic control of wind-blown snow distributions in distributed and aggregated simulations for an Arctic tundra basin, J. Hydrometeorol., 5, 735-744, 2004.

Frei, A., Tedesco, M., Lee, S., Foster, J., Hall, D. K., Kelly, R., and Robinson, D. A.: A review of global satellite-derived snow products, Adv. Space Res., 50, 1007-1029, https://doi.org/10.1016/j.asr.2011.12.021, 2012.

Hall, D. K., Riggs, G. A., and Salomonson, V. V.: Development of methods for mapping global snow cover using Moderate Resolution Imaging Spectroradiometer data, Remote Sens. Environ., 54, 127-140, https://doi.org/10.1016/0034-4257(95)00137-P, 1995.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., and Bayr, K. J.: MODIS snow-cover products, Remote Sens. Environ., 83, 181-194, https://doi.org/10.1016/S0034-4257(02)00095-0, 2002.

Hall, D. K. and Riggs, G. A.: Accuracy assessment of the MODIS snow products, Hydrol. Process., 21, 1534-1547, https://doi.org/10.1002/hyp.6715, 2007.

Hou, J. and Huang, C.: Improving mountainous snow cover fraction mapping via artificial neural networks combined with MODIS and ancillary topographic data, IEEE Trans. Geosci. Remote Sens., 52, 5601-5611, https://doi.org/10.1109/TGRS.2013.2290996, 2014.

Hou, J., Huang, C., and Zhang, Y.: ChinaAI-FSC: A Comprehensive AI-Ready MODIS Fractional Snow Cover Dataset for China (2000-2022) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17707386.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E.: Hole-filled seamless SRTM data V4, Int. Cent. Trop. Agric. (CIAT), available at: https://srtm.csi.cgiar.org (last access: 20 October 2025), 2008.

Kidwai-Khan, F., Wang, R., Skanderson, M., Brandt, C. A., Fodeh, S., and Womack, J. A.: A roadmap to artificial intelligence (AI): Methods for designing and building AI-ready data to promote fairness, J. Biomed. Inform., 154, 104654, https://doi.org/10.1016/j.jbi.2024.104654, 2024.

Klein, A. G., Hall, D. K., and Riggs, G. A.: Improving snow cover mapping in forests through the use of a canopy reflectance model, Hydrol. Process., 12, 1723-1744, https://doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1723::AID-HYP691>3.0.CO;2-2, 1998.

Klein, A. G. and Barnett, A. C.:Validation of daily MODIS snow cover maps of the Upper Rio Grande River Basin, Remote Sensing of Environment, 86, 162–176, https://doi.org/10.1016/S0034-4257(03)00097-X, 2003.

Kuter, S., Akyurek, Z., and Weber, G. W.: Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines, Remote Sens. Environ., 205, 236-252, https://doi.org/10.1016/j.rse.2017.11.021, 2018.

Kuter, S.: Completing the machine learning saga in fractional snow cover estimation from MODIS Terra reflectance data: Random forests versus support vector regression, Remote Sens. Environ., 255, 112294, https://doi.org/10.1016/j.rse.2021.112294, 2021.

Kuter, S., Bolat, K., and Akyurek, Z.: A machine learning-based accuracy enhancement on EUMETSAT H-SAF H35 effective snow-covered area product, Remote Sens. Environ., 272, 112947, https://doi.org/10.1016/j.rse.2022.112947, 2022.

Liang, X., Liu, Q., Wang, J., Chen, S., and Gong, P.: Global 500 m seamless dataset (2000-2022) of land surface reflectance generated from MODIS products, Earth Syst. Sci. Data., 16, 177-200, https://doi.org/10.5194/essd-16-177-2024, 2024.

Liu, X., Kan, X., Zhang, Y., Zhu, L., Liu, Q., Zhou, Z., and Ma, G.: FSC-USNet: Fractional snow cover retrieval on the Tibetan Plateau by integrating improved attention mechanisms, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 17, 10083-10096, https://doi.org/10.1109/JSTARS.2024.3360087, 2024.

Liu, Y., Liu, R., Chen, J., Wei, X., Qi, L., and Zhao, L.: A global annual fractional tree cover dataset during 2000-2021 generated from realigned MODIS seasonal data, Sci. Data, 11, 832, https://doi.org/10.1038/s41597-024-03671-9, 2024.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., and Gascon, F.: Sen2Cor for Sentinel-2, Proc. SPIE, 10427, 37-48, https://doi.org/10.1117/12.2278218, 2017.

Markham, B. L., Storey, J. C., Williams, D. L., and Irons, J. R.: Landsat sensor performance: history and current status, IEEE Trans. Geosci. Remote Sens., 42, 2691-2694, https://doi.org/10.1109/TGRS.2004.840720, 2004.

Metsämäki, S., Mattila, O. P., Pulliainen, J., Niemi, K., Luojus, K., and Böttcher, K.: An optical reflectance model-based method for fractional snow cover mapping applicable to continental scale, Remote Sens. Environ., 123, 508-521, https://doi.org/10.1016/j.rse.2012.04.010, 2012.

Mudryk, L., Santolaria-Otín, M., Krinner, G., Ménégoz, M., Derksen, C., Brutel-Vuilmet, C., Brady, M., and Essery, R.: Historical Northern Hemisphere snow cover trends and projected changes in the CMIP6 multi-model ensemble, The Cryosphere, 14, 2495-2514, https://doi.org/10.5194/tc-14-2495-2020, 2020.

Painter, T. H., Dozier, J., Roberts, D. A., Davis, R. E., and Green, R. O.: Retrieval of subpixel snow-covered area and grain size from imaging spectrometer data, Remote Sens. Environ., 85, 64-77, https://doi.org/10.1016/S0034-4257(02)00187-6, 2003.

Painter, T. H., Rittger, K., McKenzie, C., Slaughter, P., Davis, R. E., and Dozier, J.: Retrieval of subpixel snow covered area, grain size, and albedo from MODIS, Remote Sens. Environ., 113, 868-879, https://doi.org/10.1016/j.rse.2009.01.001, 2009.

Pan, F., Jiang, L., Wang, G., Pan, J., Huang, J., Zhang, C., Cui, H., Yang, J., Zheng, Z., Wu, S., and Shi, J.: MODIS daily cloud-gap-filled fractional snow cover dataset of the Asian Water Tower region (2000-2022), Earth Syst. Sci. Data, 16, 2501--2523, https://doi.org/10.5194/essd-16-2501-2024, 2024.

Poduval, B., McPherron, R. L., Walker, R., Himes, M. D., Pitman, K. M., Azari, A. R., Shneider, C., Tiwari, A. K., Kapali, S., Bruno, G., and Georgoulis, M. K.: AI-ready data in space science and solar physics: problems, mitigation and action plan, Front. Astron. Space Sci., 10, 1203598, https://doi.org/10.3389/fspas.2023.1203598, 2023.

Qiu, S., Zhu, Z., Shang, R., and Crawford, C. J.: Can Landsat 7 preserve its science capability with a drifting orbit?, Sci. Remote Sens., 4, 100026, https://doi.org/10.1016/j.srs.2021.100026, 2021.

Raleigh, M. S., Rittger, K., Moore, C. E., Henn, B., Lutz, J. A., and Lundquist, J. D.: Ground-based testing of MODIS fractional snow cover in subalpine meadows and forests of the Sierra Nevada, Remote Sens. Environ., 128, 44-57, https://doi.org/10.1016/j.rse.2012.09.016, 2013.

Rittger, K., Painter, T. H., and Dozier, J.: Assessment of methods for mapping snow cover from MODIS, Adv. Water Resour., 51, 367-380, https://doi.org/10.1016/j.advwatres.2012.03.002, 2013.

Salomonson, V. V. and Appel, I.: Estimating fractional snow cover from MODIS using the normalized difference snow index, Remote Sens. Environ., 89, 351-360, https://doi.org/10.1016/j.rse.2003.10.016, 2004.

Salomonson, V. V. and Appel, I.: Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results, IEEE Trans. Geosci. Remote Sens., 44, 1747-1756, https://doi.org/10.1109/TGRS.2006.876029, 2006.

Stillinger, T., Rittger, K., Raleigh, M. S., Michell, A., Davis, R. E., and Bair, E. H.: Landsat, MODIS, and VIIRS snow cover mapping algorithm performance as validated by airborne lidar datasets, The Cryosphere, 17, 567-590, https://doi.org/10.5194/tc-17-567-2023, 2023.

Tan, X., Wu, Z., Mu, X., Gao, P., Zhao, G., Sun, W., and Gu, C.: Spatiotemporal changes in snow cover over China during 1960-2013, Atmos. Res., 218, 183-194, https://doi.org/10.1016/j.atmosres.2018.11.018, 2019.

Tang, W., Zhou, J., Ma, J., Wang, Z., Ding, L., Zhang, X., and Zhang, X.: TRIMS LST: a daily 1 km all-weather land surface temperature dataset for China's landmass and surrounding areas (2000-2022), Earth Syst. Sci. Data, 16, 387-419, https://doi.org/10.5194/essd-16-387-2024, 2024.

Thackeray, C. W. and Fletcher, C. G.: Snow albedo feedback: Current knowledge, importance, outstanding issues and future directions, Prog. Phys. Geogr., 40, 392-408, https://doi.org/10.1177/0309133315620999, 2016.

U.S. National Science Foundation: Dear Colleague Letter: National Artificial Intelligence Research Resource (NAIRR) Pilot seeks datasets to facilitate AI education and researcher skill development (DCL NSF 24-093), National Science Foundation, available at: https://new.nsf.gov/funding/information/dcl-national-ai-research-resource-nairr-pilot-seeks-datasets (last access: 20 October 2025), 2024.

Xiao, X., He, T., Liang, S., Liu, X., Ma, Y., Liang, S., and Chen, X.: Estimating fractional snow cover in vegetated environments using MODIS surface reflectance data, Int. J. Appl. Earth Obs. Geoinf., 114, 103030, https://doi.org/10.1016/j.jag.2022.103030, 2022.

Xin, Q., Woodcock, C. E., Liu, J., Tan, B., Melloh, R. A., and Davis, R. E.: View angle effects on MODIS snow mapping in forests, Remote Sens. Environ., 118, 50-59, https://doi.org/10.1016/j.rse.2011.10.029, 2012.

Zhang, H., Zhang, F., Zhang, G., Che, T., Yan, W., Ye, M., and Ma, N.: Ground-based evaluation of MODIS snow cover product V6 across China: Implications for the selection of NDSI threshold, Sci. Total Environ., 651, 2712–2726, https://doi.org/10.1016/j.scitotenv.2018.10.111, 2019.

Zhao, Q., Hao, X., Che, T., Shao, D., Ji, W., Luo, S., Huang, G., Feng, T., Dong, L., Sun, X., and Li, H.: Estimating AVHRR snow cover fraction by coupling physical constraints into a deep learning framework, ISPRS J. Photogramm. Remote Sens., 218, 120-135, https://doi.org/10.1016/j.isprsjprs.2024.08.015, 2024.