



Benchmarking Snow Products under Imperfect Observational Constraints: The Case of ERA5-Land across Canada and Alaska

Ali Nazemi¹, Robert Sarpong^{1,2}

¹ Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, Canada

² Urban Systems Ltd., Kelowna, British Columbia, Canada

Correspondence to: Ali Nazemi (ali.nazemi@concordia.ca)

Abstract. Gridded snow products enable assessment where observations are sparse, but benchmarking their fidelity is complicated because the snow state is imperfectly observed. We formulate benchmarking as an evidence-based inference problem using ERA5-Land Snow Cover Fraction (SCF), snow density (Rho), Snow Depth (SD), and Snow Water Equivalent (SWE) across Canada and Alaska. ERA5-Land and competing products are harmonized to 25 km and evaluated against satellite retrievals and in-situ observations at daily and monthly scales. Fidelity is assessed using goodness-of-fit measures for time-series reconstruction and relative differences in long-term mean, trend, and detrended interannual variability. Fidelity varies with terrain, measurement technique, and observational density. ERA5-Land generally represents SCF and Rho more faithfully than SD and SWE, for which ERA5L consistently exhibits lower time-series fidelity relative to the competing gridded representations. Yet, ERA5-Land can be competitive, or preferred, for long-term mean SWE and trends of SD and SWE. Long-term mean and variability generally covary with time-series fidelity, whereas trend fidelity is weakly related and can exhibit trade-offs. This diagnostic dependence produces diagnostic inconsistency, with consequential outcomes such as reversing the inferred direction of regional SCF change in the Northern Arctic. Incomplete observational support creates distinct limits. For example, in the Marine West Coast Forest, gridded representations cannot be distinguished in their fidelity for SD trend where observations exist, yet yield opposing trends across unconstrained areas. In the ~800,000 km² Southern Arctic, ambiguity arises because observations are absent, while four SWE products support markedly different trend magnitudes and signs. Benchmarking should therefore assess product fidelity alongside the strength of evidence supporting it, evaluate diagnostic sensitivity, and retain unresolved disagreement where observations are absent or cannot discriminate among alternatives.

1 Introduction

Seasonal snow is a key land feature at high latitudes and elevations (Sexstone et al., 2016; Stiegler et al., 2016; Mitterer and Schweizer, 2013). It regulates surface albedo (Lackner et al., 2022; Willeit and Ganopolski, 2018), insulates the land surface (Kim et al., 2018; You et al., 2014), and plays a central role in regional water and energy cycles (Bair et al., 2016; Deng et al., 2019). From a hydrological perspective, Snow Cover Fraction, Snow Depth, snow density, and Snow Water Equivalent (hereafter SCF, SD, Rho, and SWE, respectively) provide complementary information on water availability associated with seasonal snow (Wang and Zheng, 2020; Li et al., 2017). SCF describes the fraction of land covered by snow, and influences



surface albedo and the spatial distribution of meltwater (Lehning, 2013; Sarpong et al., 2026). SD characterizes the vertical thickness of the snowpack and is relevant to snow accumulation and loss dynamics, and snow-related hazards (Shijin et al., 2022; Kazama et al., 2008). Rho physically links SD and SWE and provides an important diagnostic of snowpack state and evolution (Gao et al., 2022). SWE quantifies the water stored in the snowpack and is central to water-resource management and flood-risk assessment (Skaugen et al., 2012; Vionnet et al., 2021). These variables vary substantially with latitude, topography, land cover, and climate (Stiegler et al., 2016; Bormann et al., 2013). Climate change is also increasingly altering snow precipitation, accumulation and ablation dynamics, as well as their contributions to regional water and energy balances (e.g., Barnett et al., 2005; Musselman et al., 2021; Fang and Leung, 2023; Ryberg, 2024).

Close monitoring of snow variables is therefore critical for characterizing snow water availability, particularly under heightened climate variability and change. However, monitoring snow over large domains remains challenging because no observational system simultaneously provides high accuracy, long records, broad coverage, and spatial support suitable for regional-to-continental assessments. In-situ measurements provide the most direct observations of snow conditions and are a primary source of information on snow variables (Stähli et al., 2004; Brown et al., 2021). Emerging in-situ archives increasingly consolidate historical observations over large domains (e.g., Vionnet et al., 2021; Mortimer and Vionnet, 2025), but monitoring networks remain sparse and uneven, particularly in high-latitude and high-elevation environments. Satellite remote sensing provides complementary synoptic information (Dietz et al., 2012; Broxton and van Leeuwen, 2020; Lievens et al., 2019; Notarnicola, 2020), although retrieval capability differs markedly among snow variables and sensor types (Brubaker et al., 2005; Appel, 2018; Hassler and Lauer, 2021). Optical sensors such as the Moderate Resolution Imaging Spectroradiometer (MODIS; Hall and Riggs, 2007; see Riggs et al., 2017; Rittger et al., 2021; and Gascoin et al., 2020 for other optical sensors) can represent SCF, but cloud cover, vegetation, and polar night constrain their availability and accuracy (Molotch et al., 2004; Hall et al., 1998; Trepte et al., 2003). Retrieval of SD, Rho, and particularly SWE is more challenging (Dong et al., 2005; Schilling et al., 2024). Passive-microwave sensors, including the Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E and AMSR2), see Kelly et al. (2003) and Du et al. (2017), provide records suitable for large-scale SD, Rho, and SWE estimation, but their coarse resolution and sensitivity to terrain, vegetation, snow depth, wetness, and microstructure can affect fidelity (Foster et al., 2005; Byun and Choi, 2014). Active Synthetic Aperture Radar (SAR) systems such as Sentinel-1 (Hoppinen et al., 2024) offer finer spatial information (Marshall and Koh, 2008; Patil et al., 2020), but important limitations remain across dry, shallow, and deep snow conditions. Light Detection and Ranging (LiDAR; Deems et al., 2013) can provide high-resolution SD estimates; however, limitations related to snow conditions, temporal sampling, or spatial coverage remain. Currently, no remote-sensing system provides all relevant snow variables with the accuracy, duration, coverage, and temporal resolution required for large-scale assessments (Tsang et al., 2022).

Gridded model-based products complement these observational limitations by providing spatially and temporally continuous representations of historical snow conditions. Reanalyses are particularly valuable because they combine numerical Earth



System modelling with historically constrained atmospheric states through data-assimilation systems to reconstruct past
65 conditions consistently in space and time (Lindsay et al., 2014; Parker, 2016). Reanalysis approaches increasingly extend to
land-surface and regional hydroclimate systems (Rasmussen et al., 2023), including dedicated snow reanalyses (Fang et al.,
2022). Their fidelity, however, remains dependent on atmospheric forcing, land-surface model structure and parameterization,
spatial resolution, and the availability and type of assimilated information. Considerable biases can therefore occur (Graham
et al., 2019; Sun et al., 2018), especially when snow observations are not directly assimilated (Mortimer and Mudryk, 2025).
70 Numerous studies have consequently benchmarked reanalysis snow fields against observations and other reference datasets
(Morice et al., 2012; Rapać et al., 2015; Orsolini et al., 2019; Schumacher et al., 2020; Baxter et al., 2024; Ma et al., 2023).
Collectively, these evaluations suggest comparatively better representation of SCF and Rho (Gao et al., 2022; Akyurek et al.,
2023), while larger discrepancies are commonly reported for SD and SWE (Kouki et al., 2023; Monteiro and Morin, 2023).

Despite these advances, benchmarking gridded snow products involves a fundamental epistemic problem, because the state
75 against which a product is evaluated is itself imperfectly known. From a broader perspective, Oreskes et al. (1994) argued that
numerical models of natural systems cannot be verified or validated in the strong sense of demonstrating their truth, because
observations and model inputs are incomplete and multiple representations may remain consistent with the available evidence.
Agreement with observations can strengthen confidence, but cannot demonstrate that a model uniquely represents the
underlying natural state. This epistemic limitation is often incompletely explored in product benchmarking (e.g., Hamm et al.,
80 2020; Xu et al., 2022; Fatolahzadeh Gheysari et al., 2024), and is particularly consequential for snow, where observations
remain sparse, uneven, and subject to scale and representativeness mismatch with gridded products. Recent cryosphere
literature has increasingly recognized this limitation. Because neither observations nor model-based products provide an
unequivocal representation of snow at all relevant scales, integration and cross-comparison among multiple evidences are
suggested (Huning and AghaKouchak, 2020), typically after harmonizing to comparable spatial and temporal support (Lei et
85 al., 2023; Kouki et al., 2023). Yet agreement among products cannot itself establish fidelity to the underlying snow state
(Mudryk et al., 2015). Mortimer et al. (2020) distinguished evaluation against independent observations from intercomparison
among nine gridded products, while Mudryk et al. (2025) extended this progression toward multi-evidence and multi-criteria
evaluation. Benchmarking gridded snow products is therefore more appropriately viewed as inference from multiple imperfect,
differently supported, and potentially non-independent representations of an incompletely observed snow state. What remains
90 less systematically addressed is how limitations in observational support translate into limits on the strength of inference itself,
including whether competing products can be discriminated, whether conclusions remain transferable to observationally
unconstrained regions, and whether different benchmarking diagnostics support consistent assessments of fidelity.

Beyond this epistemic limitation, a related inferential problem concerns what aspect of fidelity is being evaluated and how.
Mortimer et al. (2020) found that no single product performed best across all benchmarking criteria. Conventional Goodness-
95 of-Fit (GoF) measures characterize correspondence between paired time series, whereas long-term mean, interannual



variability, trend, persistence, and spatial organization represent distinct characteristics. Mudryk et al. (2015) showed that relatively strong correspondence in temporal anomalies can coexist with substantial differences in climatological snow mass. Thus, good time-series correspondence does not necessarily imply faithful reconstruction of application-relevant characteristics. Likewise, reproducing the broad spatial organization of a snow field, a common focus of gridded-product evaluations (Henn et al., 2018; Monteiro and Morin, 2023; Scherrer et al., 2023), does not imply spatially unstructured discrepancies; errors may vary systematically across space or with characteristics of the available observations.

Here, we formulate benchmarking as a joint epistemic and diagnostic inference problem. Two conditions are central to this formulation. First, different fidelity dimensions or performance metrics can lead to different inferences about relative fidelity. Second, observational support varies in availability, duration, measurement footprint, and spatial representativeness. Together, these conditions create three epistemic limits on inference. First, diagnostic inconsistency may arise when conventional GoF measures and the ability to reproduce specific statistical characteristics lead to different assessments of relative fidelity. Second, non-identifiability occurs when competing gridded products cannot be meaningfully distinguished where observations exist but provide substantially different estimates where observations are absent. Third, ambiguity arises where no observational constraint is available to discriminate among alternative gridded representations. These distinctions separate product fidelity from the strength with which that fidelity can be inferred. Accordingly, benchmarking is not used here to establish a gridded product as a true representation of snow variable(s), but to quantify its correspondence with available evidence and determine the strength and limits of the conclusions that can be drawn from that correspondence.

We demonstrate this framework using SCF, SD, Rho, and SWE from the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) land reanalysis (ERA5-Land; Muñoz-Sabater et al., 2021), across Canada and Alaska. ERA5-Land provides a particularly informative test case because previous benchmarking efforts have reached contrasting conclusions regarding the fidelity of its snow fields. Several regional evaluations have reported substantial biases in ERA5 and ERA5-Land snow fields, particularly in mountainous environments (Orsolini et al., 2019; Kouki et al., 2023; Monteiro and Morin, 2023; Akyurek et al., 2023). In contrast, Mudryk et al. (2025) ranked ERA5-Land first among 23 Northern Hemisphere gridded SWE products based on cumulative performance across SWE climatology, variability, and trends, with particularly strong performance over mountainous North America. Rather than necessarily being contradictory, these conclusions may reflect differences in reference data, evaluation criteria, spatial support, and observational constraint. We specifically address four questions: (1) How well does ERA5-Land reproduce observed snow variables, and how does its performance compare with competing gridded products? (2) How does inferred fidelity depend on spatial gradients and attributes of observational support? (3) Does fidelity according to conventional GoF measures imply fidelity in reconstructing snow characteristics? (4) Where do available observations permit gridded representations to be discriminated, and where do non-identifiability or ambiguity limit inference about their relative fidelity? Together, these questions use ERA5-Land as a test case for the broader problem of benchmarking gridded snow products under imperfect observational constraints.



2 Case study and data support

2.1 Study domain and available data

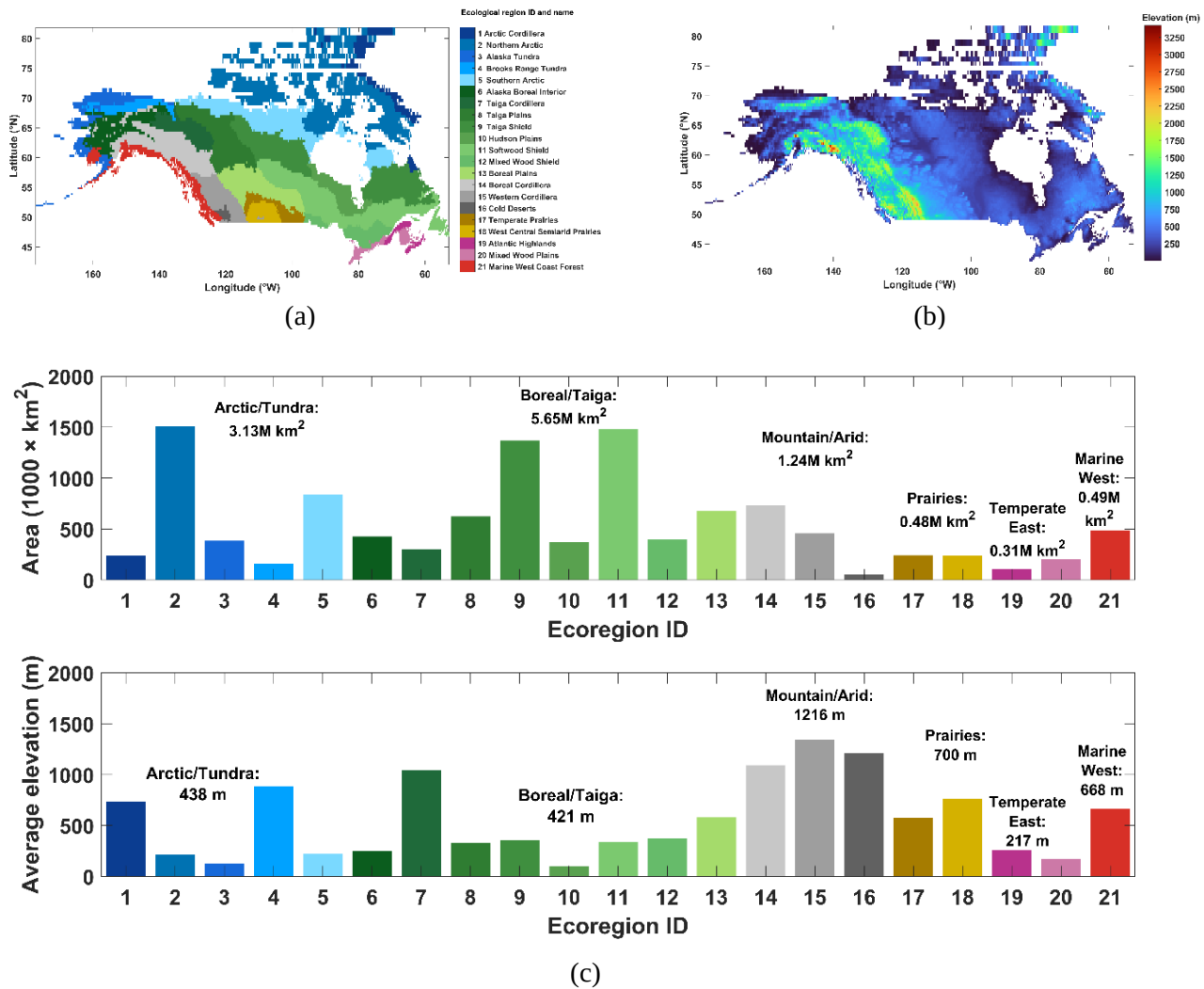
130 Canada and Alaska are selected as the benchmarking domain because together they form a large, contiguous snow-dominated region of ~12 million km² (~8% of global land area) in which seasonal snow strongly influences regional climate, ecosystems, water availability, and socio-economic activities (Boelman et al., 2019; Callaghan et al., 2011; DeBeer et al., 2016). Snow conditions vary markedly across this domain in response to strong gradients in elevation, land cover, continentality, proximity to large water bodies, and climate variability and change (Bitz and Battisti, 1999; Brown, 2010; Bormann et al., 2013; Nazemi
 135 et al., 2017). Observational support is similarly heterogeneous, ranging from relatively well-monitored southern and lower-elevation areas to sparsely observed northern and mountainous environments, making the domain particularly suitable for examining how observational support affects benchmarking conclusions.

Benchmarking is organized across 21 ecological regions that partition northern North America into areas with broadly similar geology, climate, vegetation, and ecosystems (Omernik, 2004; Wright et al., 1998; McMahon et al., 2004), thereby providing
 140 a consistent spatial framework for regional comparison (Omernik and Griffith, 2014). The ecological regions are (1) Arctic Cordillera, (2) Northern Arctic, (3) Alaska Tundra, (4) Brooks Range Tundra, (5) Southern Arctic, (6) Alaska Boreal Interior, (7) Taiga Cordillera, (8) Taiga Plains, (9) Taiga Shield, (10) Hudson Plains, (11) Softwood Shield, (12) Mixed Wood Shield, (13) Boreal Plains, (14) Boreal Cordillera, (15) Western Cordillera, (16) Cold Deserts, (17) Temperate Prairies, (18) West Central Semiarid Prairies, (19) Atlantic Highlands, (20) Mixed Wood Plains, and (21) Marine West Coast Forest. Figure 1(a)
 145 groups these regions into six eco-clusters: Arctic/Tundra (regions 1-5), Boreal/Taiga (6-13), Mountainous/Arid (14-16), Prairies (17-18), Temperate East (19-20), and Marine West (21). These numerical identifiers and colour codes are used throughout the paper to refer to ecological regions and eco-clusters. Figure 1(b) shows the 25 km mean elevation distribution across the domain, ranging from approximately 0 to 3400 m, while Fig. 1(c) summarizes area and mean elevation by ecological region and eco-cluster. Northern Arctic is the largest region (>1.5 million km²), whereas Cold Deserts is the smallest (~50,000
 150 km²); Hudson Plains and Western Cordillera respectively have the lowest (<100 m) and highest (~1350 m) mean elevations among ecological regions.

The benchmarking draws on four categories of data support, i.e., the subject product, competing gridded representations, observational constraints, and elevation data (Table 1). ERA5-Land (hereafter ERA5L) is the subject product and provides SCF, Rho, SD, and SWE at hourly resolution; these fields are aggregated to daily and monthly scales (see Section 3.1 below).
 155 At the daily scale, ERA5L SD is compared with CMC daily SD. At the monthly scale, ECCC provides a competing gridded representation for ERA5L SCF; CMC provides competing representations for Rho and SD; and CMC, ECCC, and unified AMSR-E/AMSR2 provide three competing representations for SWE. These competing products for SWE span assimilated-analysis, blended multi-product, and satellite-retrieval approaches. MODIS (Hall et al., 2002) provides the observational constraint for monthly SCF, whereas daily in-situ observations available from the Northern Hemisphere in-situ Snow Water



160 Equivalent dataset (NorSWE; Mortimer and Vionnet, 2025) provide in-situ constraints for Rho, SD, and SWE fields of ERA5L and competing gridded representations. Because snow conditions vary strongly with elevation, we also use the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010; Danielson and Gesch, 2011) to characterize the elevation dependence of fidelity and discrepancies in ERA5L and the competing gridded representations. These datasets are described below.



165 **Figure 1: Study domain and spatial framework using ecological regions: (a) spatial extent of the 21 ecological regions and six eco-clusters across Canada and Alaska; (b) mean elevation from GMTED2010 (Danielson and Gesch, 2011) aggregated to the 25 km grid; and (c) area (1000 km²; top) and mean elevation (m; bottom) across ecological regions and eco-clusters.**



Table 1. Datasets used in benchmarking, including their role, product type, supported snow variable(s), native temporal and spatial scales, and period of data record.

Role in benchmarking	Product and type	Variable	Temporal scale	Spatial scale	Available data period
Subject product	ERA5-Land (ERA5L): Offline land reanalysis	Rho	Hourly	~9 km ($0.1^\circ \times 0.1^\circ$)	1950 –present
		SD			
		SWE			
		SCF			
Competing gridded representations	CMC: Assimilated snow analysis	SD	Daily	~24 km	1998 – 2020
		SWE	Monthly		
		Rho (derived)	Monthly		
	ECCC: Blended multi-product snow dataset	SWE	Monthly	~25 km ($0.25^\circ \times 0.25^\circ$)	1981 – 2016
		SCF			
	AMSR-E/AMSR2: Satellite SWE retrieval	SWE	5-day	25 km	2002 –present (9-month gap in 2011-2012)
Observational constraints	MODIS: Satellite SCF retrieval	SCF	Monthly	~5 km ($0.05^\circ \times 0.05^\circ$)	2000 –present
	NorSWE: In-situ observational archive	Rho (derived)	Daily	Point, transect, flight-line	1979 – 2021
		SD			
		SWE			
Elevation support	GMTED2010: Digital Elevation Model	Mean elevation	N/A	~1 km	N/A

2.2 ERA5-Land snow fields

ERA5L provides hourly snow fields at approximately 9 km (0.1°) resolution from 1 January 1950 to present. It is produced through an offline integration of the Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL; Balsamo et al., 2009), forced by downscaled ERA5 atmospheric fields (Hersbach et al., 2020). ERA5L does not directly assimilate land or snow observations during this offline integration, although its atmospheric forcing inherits the observational constraints incorporated into ERA5. Section S1 of the Supplement briefly summarizes the HTESSEL representation of Rho, SD, SWE, and SCF. The ERA5L snow fields are obtained through the ECMWF Climate Data Store Application Programming



Interface (CDS-API; see Data availability) and constitute the subject product evaluated against the observational constraints and competing gridded representations described below.

180 2.3 Competing gridded representations

The Canadian Meteorological Centre (CMC) snow analysis (Brown et al., 2003) has been widely used as a reference or comparison dataset in North American snow studies (Frei et al., 2005; Niu and Yang, 2007; Brown and Frei, 2007; Cooper et al., 2018; Yoon et al., 2022), but is treated here as a competing gridded representation for SWE and SD, from which monthly Rho can be derived. The product provides daily analysed SD and monthly estimated SWE over the Northern Hemisphere at approximately 24 km resolution for 1998-2020. Daily SD combines in-situ snow-depth observations with a first-guess SD field generated by a temperature-index snowpack model forced with analysed temperature and forecast precipitation. The in-situ observations include surface synoptic SD reports and other observations distributed through the World Meteorological Organization information system. Monthly SWE is estimated from monthly mean analysed SD together with monthly climatological Rho estimates (Sturm et al., 2010). Section S2 of the Supplement summarizes the analysis procedure. Daily SD and monthly SWE are obtained from the NASA National Snow and Ice Data Center (NSIDC; Brown and Brasnett, 2010; see Data availability). Monthly SWE is unavailable during July-September, and grid cells flagged by the product developers for unrealistically high SD or SWE are excluded. Monthly CMC Rho is subsequently derived from the harmonized SWE and SD fields as described below in Section 3.1.

The Environment and Climate Change Canada (ECCC) blended snow dataset provides monthly SCF and SWE over Canada and Alaska at $0.25^\circ \times 0.25^\circ$ resolution for 1981-2016 (see Data availability). Following Mudryk et al. (2015), the product blends four daily SWE estimates derived from distinct but partly dependent modelling, reanalysis, and retrieval approaches. These estimates are from (1) the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2; Global Modeling and Assimilation Office, 2016; Gelaro et al., 2017); (2) the temperature-index snow model of Brown et al. (2003), here forced by ERA-Interim temperature and precipitation without direct snow assimilation; (3) simulations from the physically based Crocus snowpack model, also forced by ERA-Interim (Brun et al., 2013); and (4) European Space Agency GlobSnow version 2, which combines passive-microwave observations, snow-emission modelling, and surface snow-depth observations to estimate SWE over non-alpine regions (Takala et al., 2011). Because GlobSnow is partly masked over mountainous terrain, its contribution is progressively reduced as mountainous fraction increases and is replaced by the mean of the other three estimates. After regridding and the mountainous-terrain adjustment, daily SWE from each source is averaged over the month, and the resulting monthly fields are combined into the multi-product mean. Monthly SCF is derived from each constituent daily SWE field by classifying a grid cell as snow covered when SWE exceeds 4 mm, averaging the resulting binary snow-presence indicator through the month, and then averaging across the four products. Accordingly, ECCC monthly SCF primarily represents the fraction of days classified as snow covered within a month rather than an instantaneous subgrid

snow-covered area fraction. Grid-cell SCF and SWE values are weighted by the fraction of land surface within each cell,
 210 excluding ocean, lakes, glaciers, and ice caps from the contributing land fraction.

As the satellite-derived competing representation of SWE, we use the AMSR-E/AMSR2 Unified Level-3 Global 5-Day 25 km
 SWE, Version 1 (Tedesco and Jeyaratnam, 2019), distributed by NSIDC (see Data availability). The record combines
 brightness-temperature observations from AMSR-E (5 June 2002 to 2 October 2011) and AMSR2 (2 July 2012 to present),
 with an inter-mission gap of approximately nine months. The native 5-day value is the maximum valid daily SWE within each
 215 5-day window rather than a 5-day mean (Tedesco and Jeyaratnam, 2019). To improve consistency between the two sensor
 records, JAXA resamples the native Level-1 brightness temperatures to unified Level-1R fields using antenna-pattern matching
 based on the Backus-Gilbert inverse formulation (Backus and Gilbert, 1967; Maeda et al., 2016; Tedesco and Jeyaratnam,
 2019). The same Northern Hemisphere SWE retrieval framework of Tedesco and Jeyaratnam (2016) is then applied to both
 mission segments. Section S3 of the Supplement summarizes the retrieval procedure. Although the unified AMSR-E/AMSR2
 220 product (hereafter AMSR for brevity) is based on measured passive-microwave radiances, SWE is retrieved rather than directly
 observed and depends on electromagnetic modelling, effective grain-size estimation, retrieval coefficients, and a modelled
 spatiotemporal snow-density scheme (Pulliainen, 2006; Sturm et al., 2010; Tedesco and Jeyaratnam, 2016). It is therefore
 treated as a satellite-derived competing representation rather than a reference dataset.

2.4 Observational constraints

225 MODIS SCF products are widely used as observation-based constraints in snowpack assessments (e.g., Saavedra et al., 2018;
 Parajka and Blöschl, 2008; Dong, 2018; Di Marco et al., 2021). We use the MODIS/Terra monthly Level-3 global snow-cover
 product as the observational constraint for ERA5L's SCF field. The monthly field is generated from daily SCF observations
 of the MODIS/Terra Daily L3 Global 0.05° Climate Modeling Grid product (MOD10C1), providing a gridded monthly
 representation suitable for large-domain assessment (Hall et al., 2002; Maurer et al., 2003). Because MODIS SCF is itself a
 230 satellite retrieval, it is treated as an observation-based constraint rather than an error-free representation of the snow state.
 Section S4 of the Supplement summarizes the retrieval and compositing procedure. Monthly MODIS SCF is obtained from
 NSIDC at approximately 5 km resolution for 1 March 2000-31 December 2023 (see Data availability).

NorSWE is a harmonized compilation of quality-controlled Rho, SD, and SWE observations spanning 1979-2021 and was
 developed specifically to support evaluation of gridded SWE products (Mortimer and Vionnet, 2025). It contains 11,593,790
 235 valid SWE observations from 10,153 sites across North America, Russia, Finland, Norway, Switzerland, and Nepal, integrating
 measurement techniques with markedly different spatial footprints and sampling frequencies. Manual gravimetric snow
 surveys, or snow courses, account for ~66% of sites but only ~8% of SWE observations; they generally average approximately
 5-15 paired SD/SWE measurements, together with additional SD measurements, along predefined transects. NorSWE also
 contains 30,889 airborne passive-gamma SWE observations from 2,357 flight lines over the United States and southern



240 Canada, which integrate substantially larger spatial footprints than point measurements. Automated observations include snow pillows, snow scales, and fixed passive-gamma sensors. Although automated instruments represent only ~11% of sites, their high sampling frequency contributes approximately 92% of all SWE observations. Where paired SWE and SD measurements are available, Rho is derived from the paired observations. NorSWE also classifies sites as mountainous or non-mountainous using a combined terrain-slope and global mountain-inventory mask and is available at daily resolution (see Data availability).

245 The strength and spatial representativeness of the NorSWE observational constraint are not uniform among sites. Point-based automated instruments generally sample areas smaller than ~100 m² and may therefore be weakly representative of multi-kilometer grid-cell snow conditions, whereas multi-point snow courses and airborne-gamma flight lines integrate substantially larger areas. Spatial coverage is especially sparse in the Arctic/Tundra eco-cluster. Mountain sites preferentially sample higher elevations but still underrepresent the highest-elevation terrain, while boreal forest and prairie environments are also

250 underrepresented. Temporal support is similarly heterogeneous, as some networks contain >30 years of consistently sampled records, whereas many sites have shorter or intermittent records, and only a small fraction of airborne-gamma flight lines satisfy long-term consistency criteria. Rather than treating all observations as equally representative of grid-scale Rho, SD, and SWE, we retain terrain setting, measurement technique, observational density, and record length as explicit attributes of observational support and examine how inferred fidelity and discrepancies vary across these attributes.

255 **2.5 Elevation data**

GMTED2010 is a global Digital Elevation Model (DEM) produced by the U.S. Geological Survey (USGS) and the National Geospatial-Intelligence Agency (NGA) by integrating multiple elevation sources (Danielson and Gesch, 2011). It provides several topographic statistics at multiple spatial resolutions and has been widely used in hydroclimatic and environmental applications (Pakoksung and Takagi, 2021; Ermolaev et al., 2017; Amatulli et al., 2018). We obtain the ~1 km mean elevation

260 field of GMTED2010 to characterize elevation across the study domain and to evaluate the elevation dependence of product fidelity and discrepancies. The data are obtained from USGS (see Data availability).

3 Methods

3.1 Harmonizing gridded data

ERA5L and competing representations differ in temporal coverage, native resolution, and grid geometry (Table 1) and are

265 therefore harmonized and paired consistently before benchmarking. We use 2 July 2002 to 31 December 2016 as the common multi-product benchmarking period, bounded by the beginning of the AMSR-E/AMSR2 record and the end of the ECCC record. The approximately nine-month AMSR inter-mission gap is retained as missing and excluded from SWE comparisons to preserve common temporal support; no values are infilled. All gridded datasets are clipped to the landmass of Canada and Alaska using the Global Administrative Areas boundaries (GADM; <https://gadm.org/data.html>). ERA5L hourly values are



270 averaged to daily values and then to monthly values on their native grid. CMC daily SD is similarly averaged to monthly values. For AMSR, monthly SWE is calculated as the unweighted arithmetic mean of all valid native 5-day values whose nominal product dates fall within a calendar month. Because each native 5-day AMSR value is itself the maximum valid daily SWE within that 5-day window, the resulting monthly AMSR series represents the mean of 5-day maxima rather than the mean of daily SWE. CMC and ECCC SWE, as well as ECCC and MODIS SCF require no temporal aggregation.

275 To allow consistent comparisons among gridded datasets, we adopt approximately 25 km as the common target spatial support and aggregate all finer-resolution datasets accordingly. ERA5L, MODIS, and GMTED2010 (Table 1) are first aggregated by area-weighted averaging to approximately 25 km. The contribution of each source cell is weighted by the fraction of its area overlapping the target cell. We apply a conservative complete-support criterion whereby, if any source cell required for an aggregation is missing, the aggregated target cell is also treated as missing. This criterion particularly reduces coverage near

280 coastlines and water bodies but ensures that retained aggregated values are based on complete contributing spatial support. Because the resulting grids remain spatially misaligned among products, all fields are subsequently remapped to the common 25 km grid used by NASA's Making Earth System Data Records for Use in Research Environments (MEaSUREs) global landscape Freeze-Thaw Earth System Data Record (FT-ESDR; Kim et al., 2017), which contains 18,060 land grid cells over the study domain. Residual grid misalignment is resolved using inverse-distance-weighted interpolation from the four nearest

285 neighbors based on Euclidean distance (Hatami and Nazemi, 2021, 2022). CMC, ECCC, and AMSR, which are already provided at approximately the target spatial resolution, are remapped directly to the same common grid. After harmonization, monthly CMC Rho is derived from coincident valid SWE and positive SD estimates.

After remapping, ERA5L is paired with each competing gridded representation on the common grid over the 2002–2016 benchmarking period. At the daily scale, only ERA5L and CMC SD can be paired (Table 1). At the monthly scale, ERA5L is

290 paired with ECCC for SCF, with CMC for Rho and SD, and with CMC, ECCC, and AMSR for SWE. Pairwise analyses retain only coincident grid and time steps with valid values in both products. For variables with multiple competing representations, comparisons are restricted to their common spatiotemporal support so that differences in inferred fidelity are not driven by unequal data coverage. Before conditioning on observations, this common support covers nearly 89%, 78%, 73%, and 84% of the domain for SCF, SD, Rho, and SWE, respectively. For consistency across variables and products, annual summaries are

295 constructed from the nine-month October–June window because CMC SWE is unavailable in July to September. Accordingly, “annual” summaries below refer to quantities calculated from the nine retained months within each calendar year.

3.2 Conditioning on observational constraints

After gridded-product pairing, the comparisons are conditioned on the corresponding observational constraints. For SCF, observational conditioning is possible only at the monthly scale (Table 1). MODIS is harmonized to the common 25 km

300 MEaSUREs grid using the same spatial procedure described above, and only grid cells with a complete MODIS record over



the retained October–June benchmarking months are retained. Because polar night prevents complete late-fall and winter retrievals at high latitudes, this criterion excludes grid cells above 64°N (Figure S1 in the Supplement), corresponding to about 42% of the study domain and including most of the Arctic/Tundra, large portions of Boreal/Taiga, and parts of Mountainous/Arid. The remaining 58% of the domain provides complete monthly MODIS support for SCF benchmarking.

305 For Rho, SD, and SWE, NorSWE provides the observational constraints. No spatial interpolation of the in-situ observations is applied; individual observations are assigned to the common 25 km grid cell containing their reported location. At the daily scale, gridded estimates are paired with observations on coincident dates. For monthly analyses, valid daily in-situ observations within each month are averaged arithmetically to form a monthly observational estimate, and gridded monthly values are paired only where such an estimate exists. These monthly values therefore represent averages of the available observations
 310 and do not necessarily constitute complete within-month means for intermittently sampled measurement types. For each variable and time scale, we first retain grid cells with cumulative observational support equivalent to at least one nine-month analysis year, defined as at least 273 daily observations or 9 monthly estimates during the benchmarking period. These grid cells are used to evaluate time-series correspondence (hereafter weak conditioning). We then identify the subset with support equivalent to at least 10 distinct analysis years, (as at least 2730 daily observations or 90 monthly estimates grouped in 10
 315 years), for evaluating reconstruction of long-term mean, trend, and interannual variability (hereafter strong conditioning). These thresholds of data availability are pragmatic choices that balance record length against spatial coverage and denote relative levels of observational support within this study rather than certainty that the underlying snow state is known.

Figures S2 and S3 in the Supplement show the grid cells satisfying these conditioning criteria at daily and monthly scales, respectively. Figures S4–S7 summarize the corresponding observational attributes, including terrain setting (non-mountainous
 320 versus mountainous), measurement technique (multi-point, single point, snow pillow, passive gamma, airborne, or mixed), observational density within each grid cell, and record length. Most selected sites contain mixed measurement types and, except under weak monthly conditioning, conditioned grid cells are concentrated in mountainous terrain (see also the elevation profiles in Figures S2 and S3). For Rho, SD, and SWE at both time scales, many grids contain more than one in-situ observation site. In such cases, we construct a grid-level observational series by taking the arithmetic mean of all valid observations
 325 available within that grid cell on the corresponding day or within the corresponding month.

Table S1 in the Supplement summarizes the number of grid cells satisfying the daily and monthly conditioning criteria for Rho, SD, and SWE across ecological regions. Observational coverage is substantially smaller for the daily analyses; for example, Rho satisfies strong conditioning in only 7 grid cells across the entire domain. Monthly SWE has the broadest in-situ support, with 832 weakly conditioned grid cells, but these still represent only about 4.6% of the 18,060-cell study domain.
 330 Coverage is particularly sparse in Arctic/Tundra regions, Hudson Plains, and West Central Semiarid Prairies. This spatially



limited support motivates the explicit analysis of observational constraint, diagnostic inconsistency, and identifiability in Sections 3.5 and the corresponding results in Section 4.4.

3.3 Benchmarking metrics

Table 2 summarizes the dimensions of the benchmarking framework. Evaluation is performed at daily and monthly scales over grid cells satisfying the corresponding observational-conditioning criteria. Two complementary dimensions of fidelity are considered: (1) time-series fidelity, which evaluates how well ERA5L and competing gridded representations reproduce observed temporal dynamics; and (2) characteristic fidelity, which evaluates how well they reconstruct the long-term mean, trend, and interannual variability of annual snow characteristics.

Table 2. Benchmarking framework for evaluating ERA5L and competing gridded representations, including fidelity dimension, diagnostic measure, observational conditioning, benchmarking subject, number of conditioned grid cells, and competing representation(s).

Benchmarking scale	Benchmarking dimension	Measure/conditioning	Benchmarking subject	Number of grids conditioned on observations	Competing gridded product(s)
Daily	Time series fidelity	PBIAS, NSE, KGE; weak conditioning	Rho time series	66	N/A
			SD time series	74	CMC
			SWE time series	122	N/A
	Characteristic fidelity	RD of long-term mean, trend, and detrended standard deviation; strong conditioning	October–June annual maximum Rho	7	N/A
			October–June annual maximum SD	30	CMC
			October–June annual maximum SWE	76	N/A
			SCF time series	10419	ECCC
Monthly	Time series fidelity	PBIAS, NSE, KGE; weak conditioning	Rho time series	667	CMC
			SD time series	723	CMC
			SWE time series	832	CMC; ECCC; AMSR
			October–June annual mean SCF	10419	ECCC
	Characteristic fidelity	RD of long-term mean, trend, and detrended standard deviation; strong conditioning	October–June annual mean Rho	41	CMC
			October–June annual mean SD	52	CMC
			October–June annual mean SWE	85	CMC; ECCC; AMSR



Time-series fidelity is quantified using three GoF measures, i.e., percent bias (PBIAS), Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), and Kling–Gupta efficiency (KGE; Gupta et al., 2009). PBIAS measures mean bias, NSE evaluates squared error relative to the observed variability, and KGE jointly assesses correlation, relative variability, and mean behavior:

$$\text{PBIAS} = \frac{\sum_{i=1}^n (X_{E/P,i} - X_{O,i})}{\sum_{i=1}^n X_{O,i}} \times 100 \quad (1)$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n (X_{E/P,i} - X_{O,i})^2}{\sum_{i=1}^n (X_{O,i} - \mu_O)^2} \quad (2)$$

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad \alpha = \frac{\sigma_{E/P}}{\sigma_O}, \quad \beta = \frac{\mu_{E/P}}{\mu_O} \quad (3)$$

where, $X_{O,i}$ denotes the observational constraint for snow field X at time i ; $X_{E/P,i}$ denotes the corresponding gridded estimate; n is the number of paired values; r is the Pearson correlation coefficient; μ_O and $\mu_{E/P}$ are the corresponding sample means; and σ_O and $\sigma_{E/P}$ are the corresponding standard deviations. Positive and negative PBIAS indicate over- and underestimation, respectively, and PBIAS = 0 indicates no aggregate bias. NSE = 1 and KGE = 1 are optimal, with progressively smaller values indicating poorer correspondence. For NSE, values below zero indicate that the mean of the observational series provides a lower squared-error benchmark than the evaluated product. GoFs are calculated over weakly conditioned grid cells (Table 2) and summarized using empirical cumulative distribution functions (ECDFs) and spatial mean values. For Rho, SD, and SWE, the spatial mean GoFs are additionally stratified by terrain setting, measurement technique, and observational density.

Characteristic fidelity is evaluated using three attributes of the annual snow series, i.e., long-term mean, trend, and interannual variability. This analysis is restricted to strongly conditioned grid cells (Table 2). At the daily scale, each calendar year is represented by the maximum daily value within the retained October–June analysis window; at the monthly scale, each year is represented by the arithmetic mean of the nine retained monthly values. The long-term mean is the average of the resulting annual series. Trend is quantified using the Theil–Sen slope estimator (Theil, 1950; Sen, 1968), providing robust inference of the gradual monotonic change. Interannual variability is quantified by the standard deviation of residuals after removing the fitted Theil–Sen trend, thereby separating year-to-year dispersion from gradual change. For each characteristic, discrepancy between a gridded representation and observational constraints can be quantified using relative difference (RD):

$$\text{RD}_j = \frac{Y_{G,j} - Y_{O,j}}{Y_O} \times 100 \quad (4)$$

In Eq. (4), RD_j is the percentage of relative difference in for statistic j ($j = 1$: long-term mean; 2: trend; 3: standard deviation); $Y_{G,j}$ is the statistic derived from ERA5L or competing gridded representation for statistic j ; and $Y_{O,j}$ is the corresponding



statistic estimated from the observational constraint. For long-term mean and standard deviation, positive and negative RD indicate over- and underestimation. For trend, the sign is interpreted relative to the signed observed slope. RD is undefined when the observational characteristic is zero and can become arbitrarily large as the denominator approaches zero. This is particularly important for trend. We therefore exclude grids for which the absolute observed annual trend is less than 0.1% of the absolute observed long-term mean per year. RDs are summarized using ECDFs and spatial mean values, and for Rho, SD, and SWE, are additionally stratified by terrain setting, measurement technique, and observational density.

3.4 Diagnosing dependencies

Spatial structure of fidelity and its association with observational length are evaluated using Kendall's tau (Kendall, 1938; Kaemo et al., 2022), a rank-based metric that avoid requiring a prescribed Gaussian residual distribution. When the analysis concerns only magnitude, $|PBIAS|$ and $|RD|$ are used. Joint dependencies between GoFs and RDs are also examined to determine whether time-series and characteristic fidelities are associated. For these joint analyses, GoFs are evaluated over the same strongly conditioned grids used for RD estimations. Because the diagnostics have different optimal directions, increasing $|PBIAS|$ and $|RD|$ indicates increasing discrepancy, whereas increasing NSE and KGE indicates improving correspondence. We consider $p\text{-value} \leq 0.05$ as an exploratory measure for statistical significance of evaluated dependencies.

3.5 Epistemic uncertainty and limits of inference

We compare the fidelity of ERA5L relative to competing gridded representations across ecological regions to examine how inference depends on spatial setting, snow variable, fidelity dimension, and diagnostic metric. We also examine how the three forms of epistemic uncertainty, i.e., diagnostic inconsistency, non-identifiability, and ambiguity, manifest and what consequences they may have for inference. For ambiguity, we quantify its magnitude for characteristic j using an empirical spread-to-signal ratio, defined as the range of estimates across the available gridded representations relative to the magnitude of their ensemble-mean estimate:

$$\mathcal{A}_j = \frac{\max_{1 \leq i \leq n} Y_{i,j} - \min_{1 \leq i \leq n} Y_{i,j}}{|\bar{Y}_j|} \quad (5)$$

where $\mathcal{A}_j$ is the ambiguity index for characteristic j , $Y_{i,j}$ is the characteristic j estimated by the representation i . n is the number of available gridded representations; and $|\bar{Y}_j|$ is the magnitude of their ensemble mean. $\mathcal{A}_j > 1$ indicates that the inter-product spread exceeds the magnitude of the ensemble-mean signal. The ambiguity index is undefined when the ensemble mean is zero and can become arbitrarily large as its magnitude approaches zero. This issue is particularly relevant to trend. Accordingly, near-zero ensemble-mean trends are excluded using an analogous screening criterion to that adopted for RD analysis. In this case, grid cells for which the magnitude of the trend is less than 0.1% of the magnitude of the long-term mean of ensemble mean per year are excluded.

4 Results

4.1 Fidelity in representing time series of snow variables

Figure 2 summarizes the fidelity of ERA5L in tracking observed daily time-series of Rho, SD, and SWE. Daily SCF cannot be evaluated because no corresponding daily observational constraint is available. Also only CMC can provide a daily competing representation for SD. From top to bottom, rows correspond to PBIAS, NSE, and KGE. Left column shows the ECDFs of GoFs. In each panel, solid vertical lines show the spatial mean values of GoFs. If a solid vertical line is missing from a panel, it means that the spatial mean is beyond the displayed range. The three middle columns show the expected GoFs across regions (mountainous vs. non-mountainous), measurement method (snow pillow, passive gamma, mixed), and observation density (1 to 8 observational sites per grid). The right column shows the significant dependencies of GoFs on longitude, latitude, elevation, and data length. Across the three GoFs, ERA5L shows consistently higher fidelity for Rho than for SD and SWE. ERA5L errors in SD and SWE are larger in non-mountainous than in mountainous grids, although substantially fewer non-mountainous grids satisfy the daily conditioning criteria (Figure S4 in the Supplement).

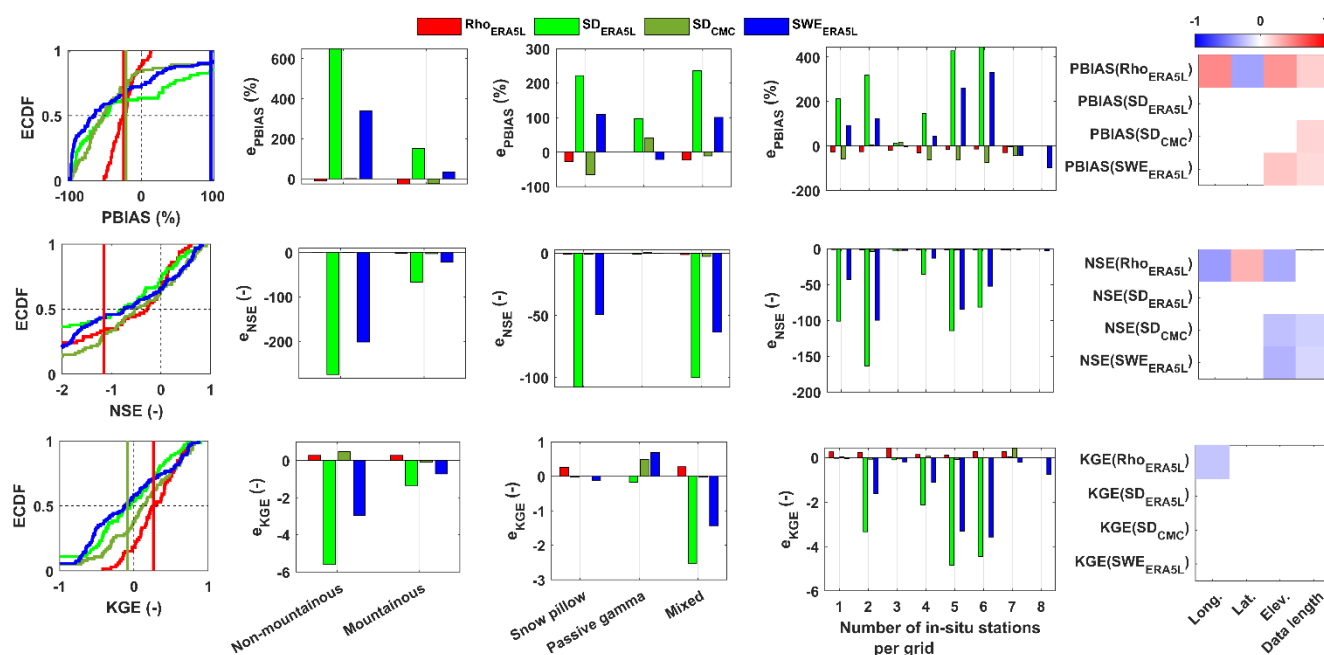


Figure 2: Assessing time-series fidelity of ERA5L for daily Rho (red), SD (ERA5L, light green; CMC, dark green), and SWE (blue). Rows show PBIAS, NSE, and KGE. From left to right, columns show ECDFs across benchmarking grid cells, stratified spatial mean of GoFs by terrain setting, measurement technique, and observational density, and significant dependencies on longitude, latitude, elevation, and observational record length (p -value ≤ 0.05). Solid vertical lines in the ECDF panels denote spatial means; dashed reference lines denote GoF = 0 and ECDF = 0.5. In the dependency panels, positive and negative associations are shown by red and blue shades, respectively.



SD and SWE performance also varies with measurement technique, with lower fidelity for mixed observations and snow pillows. Sensitivity to measurement technique and observational density is less pronounced for ERA5L Rho and CMC SD. Overall, CMC consistently outperforms ERA5L for daily SD across the three GoFs and the examined observational attributes. Higher overall GoF performance, however, does not imply spatially unstructured error. Despite its larger overall SD errors, ERA5L shows no significant dependence of daily SD performance on longitude, latitude, elevation, or record length. In contrast, CMC SD shows a significant decline in NSE with elevation, while increasing record length is associated with larger |PBIAS| and lower NSE. ERA5L Rho, although showing better overall fidelity than SD and SWE, exhibits significant dependence on at least one spatial attribute for each GoF. KGE generally shows fewer significant dependencies on spatial attributes and record length than PBIAS and NSE.

Figures 3 and 4 and Figures S8–S11 in the Supplement summarize monthly time-series fidelity. At this scale, SCF can also be conditioned on MODIS, and every ERA5L snow variable can be compared with at least one competing gridded representation. For Rho, ERA5L has a negative mean PBIAS whereas CMC is close to unbiased; however, ERA5L has higher mean NSE and KGE. For SD, ERA5L has a large positive mean PBIAS, while CMC has a smaller negative bias; CMC also has higher NSE and KGE. For SWE, ERA5L has a large positive bias whereas CMC, ECCC, and AMSR have negative mean biases. NSE for ERA5L is below zero in 75% of benchmarking grid cells, compared with 52%, 69%, and 90% for ECCC, CMC, and AMSR, respectively. ERA5L also has a negative mean KGE that is lower than AMSR, while ECCC and CMC have positive mean KGE, with ECCC showing the highest value.

Figures S8–S10 disaggregate these monthly results by terrain setting, measurement technique, and observational density. For Rho, CMC has a smaller mean |PBIAS| than ERA5L in both mountainous and non-mountainous grid cells, whereas ERA5L has higher mean NSE and KGE in non-mountainous cells. For SD, CMC has higher fidelity than ERA5L across all three GoFs and both terrain settings, although the difference is smaller in non-mountainous cells. Competing representations also generally outperform ERA5L for SWE, with some measurement-specific exceptions. ERA5L, for example, has more favorable PBIAS and KGE than CMC, ECCC, and AMSR against passive-gamma observations. Across observational-density classes, competing representations generally show higher fidelity for SD and SWE, except for PBIAS at the highest observational densities. For Rho, CMC retains the smaller |PBIAS| across all density classes, whereas ERA5L has higher NSE and KGE for cells containing 2, 3, or 6 in-situ stations; together these classes account for nearly 75% of the evaluated grid cells (Figure S5).

Figure 4 shows that monthly PBIAS and NSE exhibit significant spatial dependence for ERA5L and several competing representations, although ERA5L SD is an exception. KGE exhibits fewer significant spatial dependencies, largely confined to selected competing representations such as CMC SD and SWE. Increasing observational record length is associated with larger |PBIAS| for the snow variables and products showing significant dependence. Longer records are associated with higher NSE for ERA5L Rho but show no significant association with ERA5L SD or SWE. The association with record length differs



among competing SWE representations: NSE increases with record length for ECCC but decreases for AMSR. KGE shows no significant association with observational record length for the evaluated ERA5L or competing fields.

For monthly SCF (Figure S11 in the Supplement), ERA5L and ECCC show broadly comparable time-series fidelity. Both have mean PBIAS near zero across the 10,419 benchmarking grid cells (Table 2), and NSE and KGE are above zero in at least 80% of cells. ERA5L has the higher mean NSE, whereas ECCC has the higher mean KGE. ERA5L PBIAS and NSE also exhibit significant spatial dependence, particularly with latitude and elevation, while no significant spatial dependence is detected for the GoFs of ECCC SCF.

A direct daily-to-monthly comparison over common grid cells (62 for Rho, 71 for SD, and 112 for SWE; see Figure S12 in the Supplement) shows contrasting effects of temporal aggregation. ERA5L Rho improves from daily to monthly resolution across all three GoFs. In contrast, ERA5L SD and SWE show lower monthly than daily fidelity, and CMC SD similarly degrades across all three GoFs from daily to monthly resolution.

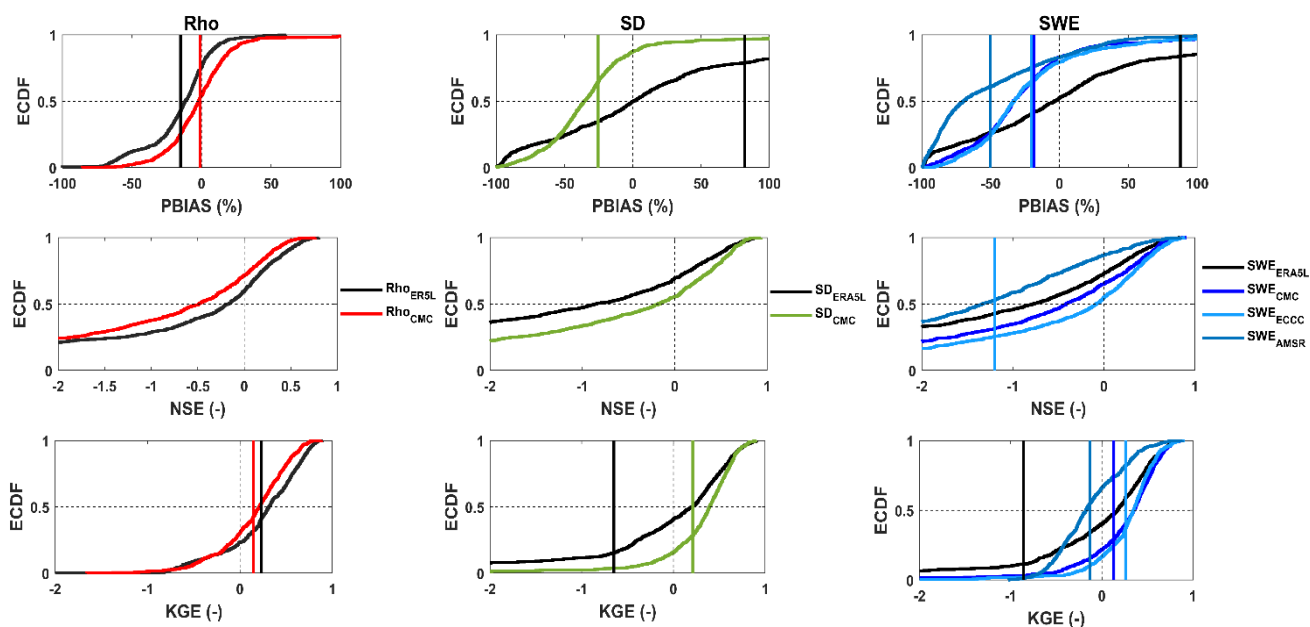


Figure 3: ECDFs of PBIAS (top row), NSE (middle row), and KGE (bottom row) for ERA5L and the corresponding competing representation(s) in reproducing monthly time series of Rho (left), SD (middle), and SWE (right). Black curves denote ERA5L and colored curves denote competing representations. Solid vertical lines indicate mean GoFs across benchmarking grid cells.

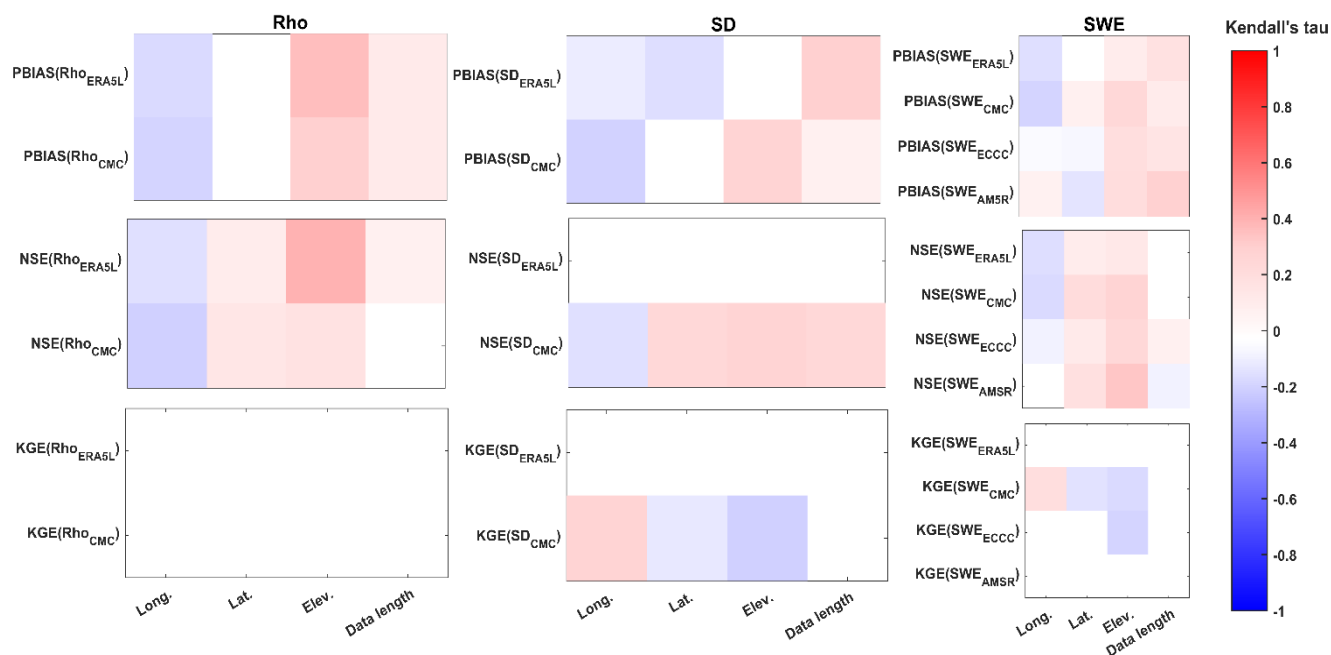


Figure 4: Significant dependencies of monthly time-series GoFs for Rho (left), SD (middle), and SWE (right) on longitude, latitude, elevation, and observational record length. Rows correspond to PBIAS (top), NSE (middle), and KGE (bottom). Only associations with p -value ≤ 0.05 are shown.

4.2 Fidelity in representing characteristics of snow variables

The fidelity of ERA5L in reconstructing characteristics of snow variables is evaluated at both daily and monthly scales. ERA5L fidelity in reconstructing SCF characteristics cannot be assessed at the daily scale because of the lack of corresponding observational constraints. We therefore evaluate the reconstruction of the long-term mean, trend, and detrended standard deviation of annual maxima of Rho, SD, and SWE relative to observational constraints (Figure 5). The configuration of Figure 5 follows that of Figure 2. The spatial-mean RD in long-term mean is slightly negative for ERA5L Rho and SWE and for CMC SD, whereas ERA5L SD shows a positive expected RD of larger magnitude. Discrepancies in long-term mean are larger in mountainous grid cells, particularly for ERA5L SD, and are generally smaller for mixed observations, again with ERA5L SD as an exception. Mean $|RD|$ also increases in grid cells with higher observational density, particularly those containing four or more in-situ stations. Significant spatial dependence in long-term-mean discrepancies is detected for ERA5L SD and SWE and for CMC SD, but not for ERA5L Rho. Discrepancies in interannual variability show broadly similar patterns to those in long-term mean. For trend, the expected RD is negative for all evaluated fields and is most negative for ERA5L SD, particularly in non-mountainous grid cells. Trend discrepancy varies little with measurement technique or observational density, although significant spatial dependence is detected for ERA5L Rho and SD.

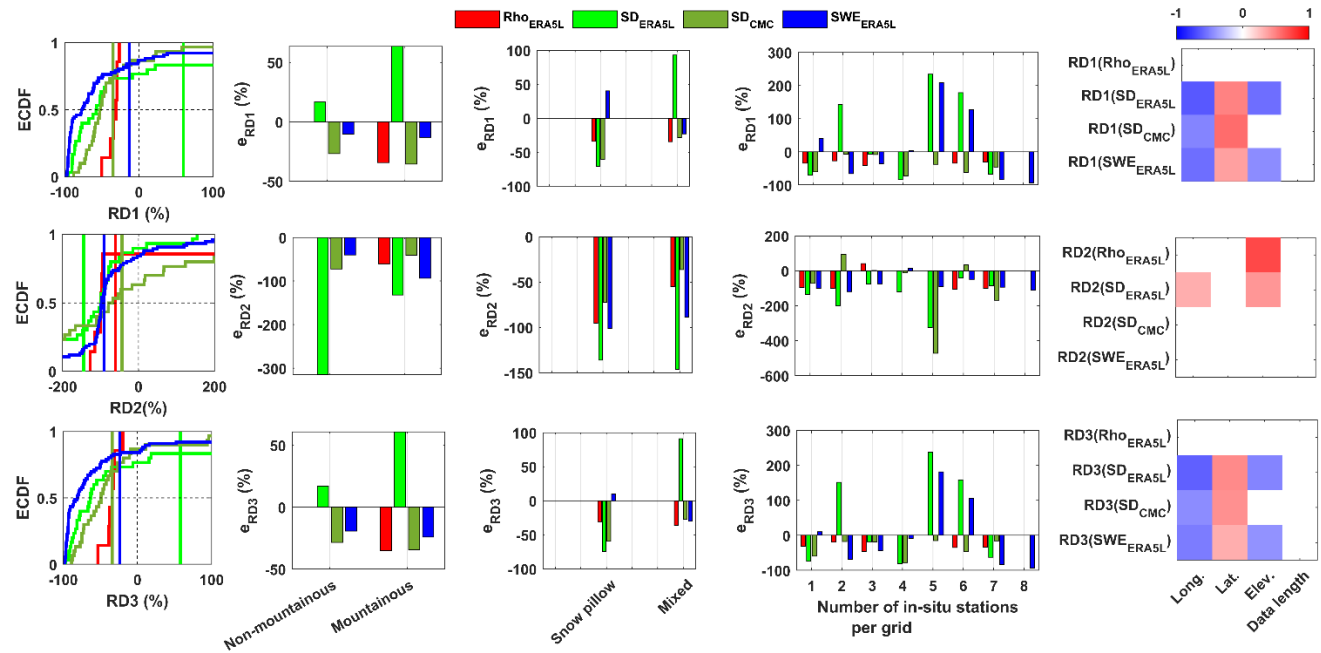


Figure 5: Assessing fidelity in reconstructing annual maxima of daily Rho (red), SD (ERA5L, light green; CMC, dark green), and SWE (blue). Rows show RD in long-term mean, trend, and detrended standard deviation. From left to right, columns show ECDFs of RDs, stratified mean RDs by terrain setting, measurement technique, and observational density, and significant dependencies of $|RD|$ on longitude, latitude, elevation, and observational record length (p -value ≤ 0.05). Solid vertical lines in ECDF panels denote spatial expected RD; dashed reference lines denote $RD = 0$ and $ECDF = 0.5$. Positive and negative dependencies are shown by red and blue shades, respectively.

Figure 6 and Figures S13–S14 in the Supplement summarize fidelity in reconstructing the long-term mean, trend, and detrended standard deviation of annual-mean values, respectively. For the long-term mean (Figure 6), ERA5L and ECCC SCF both have expected RDs near zero, although ERA5L discrepancies exhibit significant spatial dependence. ERA5L and CMC both have slightly negative expected RDs for Rho, with ERA5L showing equal or larger discrepancy magnitudes across terrain settings, measurement techniques, and most observational-density classes. Both products exhibit significant spatial dependence in discrepancies of Rho's long-term mean, while longer observational records are associated with smaller $|RD|$ for CMC. For SD, CMC has a small negative expected RD, whereas ERA5L has a large positive expected RD and substantially larger $|RD|$ across terrain settings, measurement techniques, and most observational-density classes. Both products also exhibit significant spatial dependence. For SWE, ERA5L has the smallest magnitude of expected RD among ERA5L, CMC, ECCC, and AMSR, particularly in mountainous grids and for mixed observations. Nevertheless, discrepancies of ERA5L SWE remain spatially structured and, unlike those of the competing representations, do not decrease significantly with increasing observational record length. Results for detrended standard deviation show broadly similar patterns (Figure S14 in the Supplement).

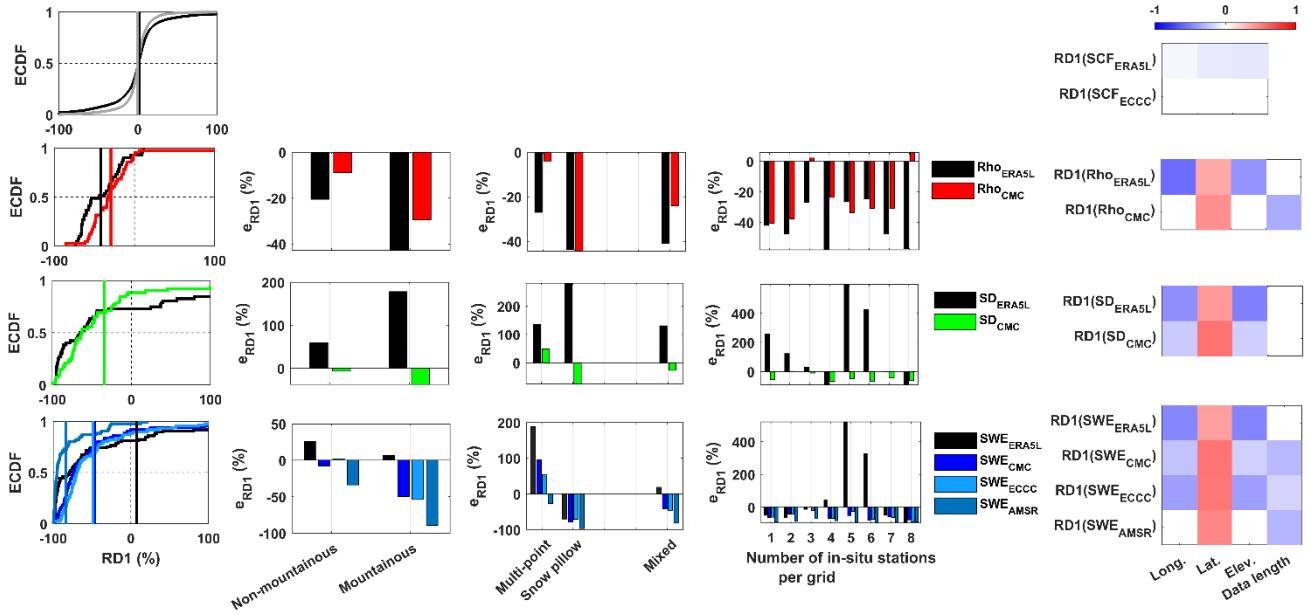


Figure 6: Fidelity in reconstructing the annual long-term mean of SCF, Rho, SD, and SWE from top to bottom, respectively. The left column shows ECDFs of RDs across benchmarking grid cells; solid vertical lines denote expected RD. The three middle columns show stratified expected RDs by terrain setting, measurement technique, and observational density, with black bars for ERA5L and coloured bars for competing representation(s). The right column shows significant dependencies of discrepancies on longitude, latitude and elevation, as well as observational record length (p -value ≤ 0.05); positive and negative associations are shown by red and blue shades, respectively.

Trend results (Figure S13 in the Supplement) are more variable across snow variables and benchmarking grids. ERA5L and ECCC both show positive expected RD in SCF trend of approximately 100%, and ERA5L trend discrepancy is spatially structured. For Rho, both ERA5L and CMC show negative expected RD below -100% , with the more negative value for ERA5L; ERA5L also has larger $|RD|$ across terrain settings, all measurement techniques except multi-point observations, and most classes of observational density. For SD and SWE, ERA5L likewise shows negative spatial-mean RD below -100% , but its discrepancy magnitude is generally smaller than that of the three competing representations, particularly in mountainous grid cells, for snow-pillow and mixed observations, and at grid cells with lower observational density. The magnitude of SD trend discrepancy decreases with elevation, whereas SWE trend discrepancy increases toward western longitudes.

4.3 Diagnostic consistency between time-series and characteristic fidelity

Associations between fidelities in time series and characteristic reconstructions can be evaluated only where strong observational conditioning is available. At the monthly scale (Figure 7), long-term mean and detrended standard deviation



generally show the expected diagnostic consistency. For ERA5L SCF, larger magnitude of PBIAS is associated with larger magnitudes of RD in reconstructing characteristics, whereas higher NSE is associated with smaller magnitudes of RD. Similar relationships occur for ERA5L Rho, SD, and SWE, with higher KGE additionally associated with smaller long-term-mean and standard-deviation discrepancies. No significant GoF–RD association is detected for the competing ECCC SCF product. For competing Rho, SD, and SWE representations, long-term mean and standard deviation generally show the same diagnostically consistent pattern as ERA5L. Trend fidelity is much less connected to time-series fidelity. KGE shows no significant association with trend discrepancy for any ERA5L snow variable, and no competing representation shows a significant GoF–RD association except AMSR SWE, for which larger magnitude of PBIAS is associated with smaller trend discrepancy. Thus, agreement between time-series and characteristic fidelity is strongest for long-term mean and interannual variability and weakest, and occasionally reversed, for trend.

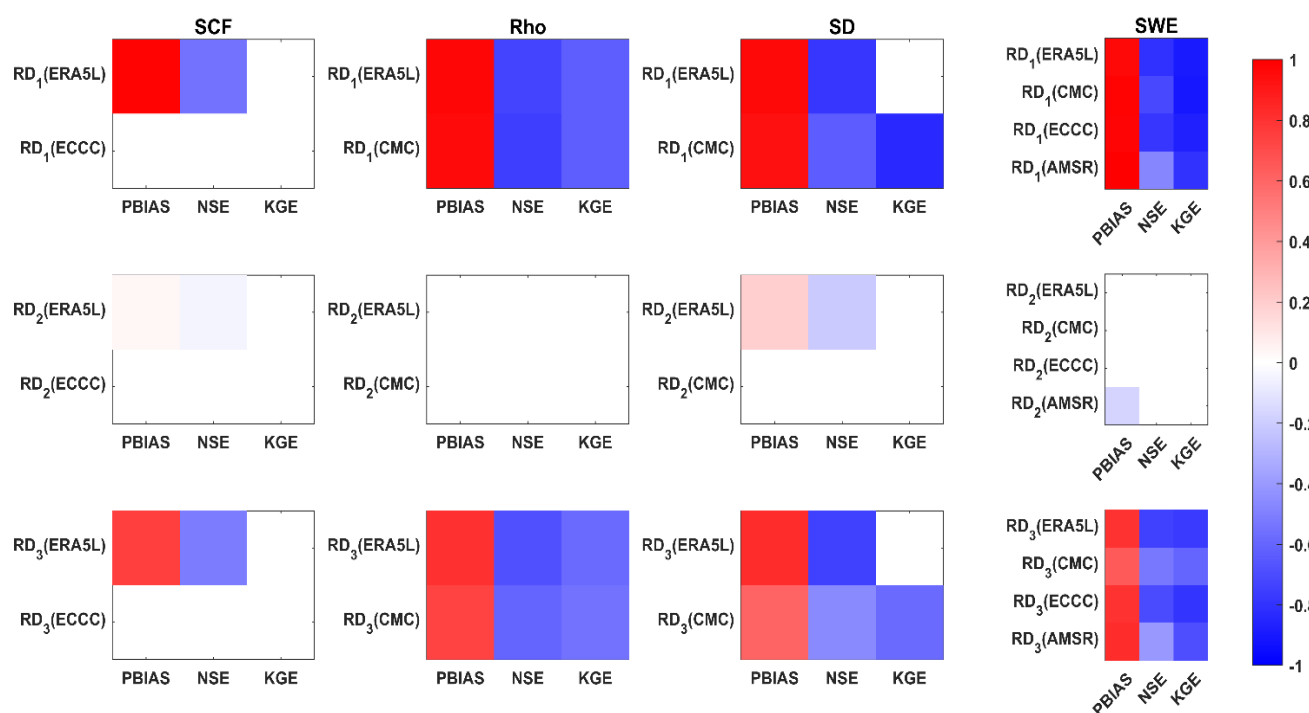


Figure 7: Significant joint associations between monthly time series GoFs and discrepancy in reconstructing October–June annual-mean snow variables. Rows correspond to RD in long-term mean (RD₁), trend (RD₂), and detrended standard deviation (RD₃); columns correspond to SCF, Rho, SD, and SWE. Associations are evaluated over the same strongly conditioned grid cells used for RD estimation, and only Kendall tau’s associations with p -value ≤ 0.05 are shown.



530 At the daily scale (Figure S15 in the Supplement), no significant GoF–RD association is detected for Rho, although only seven
 grid cells are available and the test therefore has limited power. For SD estimates of ERA5L and CMC, larger magnitude of
 PBIAS is associated with larger magnitude of RD in long-term mean and detrended standard deviation, whereas higher NSE
 and KGE are associated with smaller magnitude of discrepancies in these two characteristics. SWE estimates of ERA5L show
 the same diagnostically consistent pattern for long-term mean and standard deviation. Trend behaves differently. No GoF is
 535 significantly associated with discrepancies in trends of SD for ERA5L or CMC, while SWE estimates of ERA5L show the
 opposite direction to that expected from overall fidelity: larger magnitude of PBIAS is associated with smaller magnitude of
 RD in trend, whereas higher NSE and KGE are associated with larger magnitude of RD in trend. This constitutes a clear
 diagnostic inconsistency between time-series fidelity and trend fidelity for annual-maximum SWE.

4.4 Limits of inference under incomplete observational constraints

540 Comparison of ERA5L and competing products against the available observational constraints allows the relative fidelity of
 alternative gridded representations to be assessed across ecological regions at daily and monthly scales. Figure S16 in the
 Supplement illustrates this assessment for daily SD. CMC shows higher relative fidelity than ERA5L across all ecological
 regions according to the GoFs and the RDs in long-term mean and detrended standard deviation. Trend reconstruction provides
 two exceptions. In the Boreal Cordillera, CMC is favoured by the GoFs under weak conditioning, whereas ERA5L is favoured
 545 by the RD in the trend of annual maximum SD under strong conditioning. The preferred representation therefore depends on
 the fidelity dimension and level of observational conditioning used to support the inference, demonstrating diagnostic
 inconsistency in the resulting ranking. In Cold Deserts, this diagnostic inconsistency co-occurs with non-identifiability. CMC
 and ERA5L cannot be distinguished according to the GoFs under weak conditioning, whereas ERA5L shows higher relative
 fidelity for trend reconstruction under strong conditioning. From a broader perspective, Figure S17 in the Supplement shows
 550 that observational conditioning for daily SD is sparse across ecological regions. The largest weakly conditioned fraction is
 only 4% of the land area of an ecological region (Western Cordillera), while 13 of the 21 ecological regions remain entirely
 unconstrained. Where observational constraints are absent, the relative fidelity of competing gridded representations cannot
 be empirically evaluated; and, differences among their estimates represent ambiguity in the state of snow condition given the
 alternative representations considered.

555 Similar inferential limitations also manifest at the monthly scale (Figure 8). Although observational support increases
 substantially relative to the daily scale, the degree of observational constraint remains limited for Rho, SD, and SWE. Even in
 the most observationally constrained case, only about 25% of grid cells within an ecological region are conditioned by
 observations (SWE in Cold Deserts), while three Arctic regions, i.e., Arctic Cordillera, Alaska Tundra, and Southern Arctic,
 remain entirely unconstrained by observations, preventing empirical assessment of the relative fidelity of competing gridded
 560 representations (see Figure S18 in the Supplement). Even for SCF, which has the broadest observational support and for which
 several ecological regions are fully conditioned by MODIS, the inferred relative fidelity varies across ecological regions and



according to the fidelity dimension and diagnostic metric considered. Consequently, the three manifestations of epistemic uncertainty, diagnostic inconsistency, non-identifiability, and ambiguity, also emerge at the monthly scale.

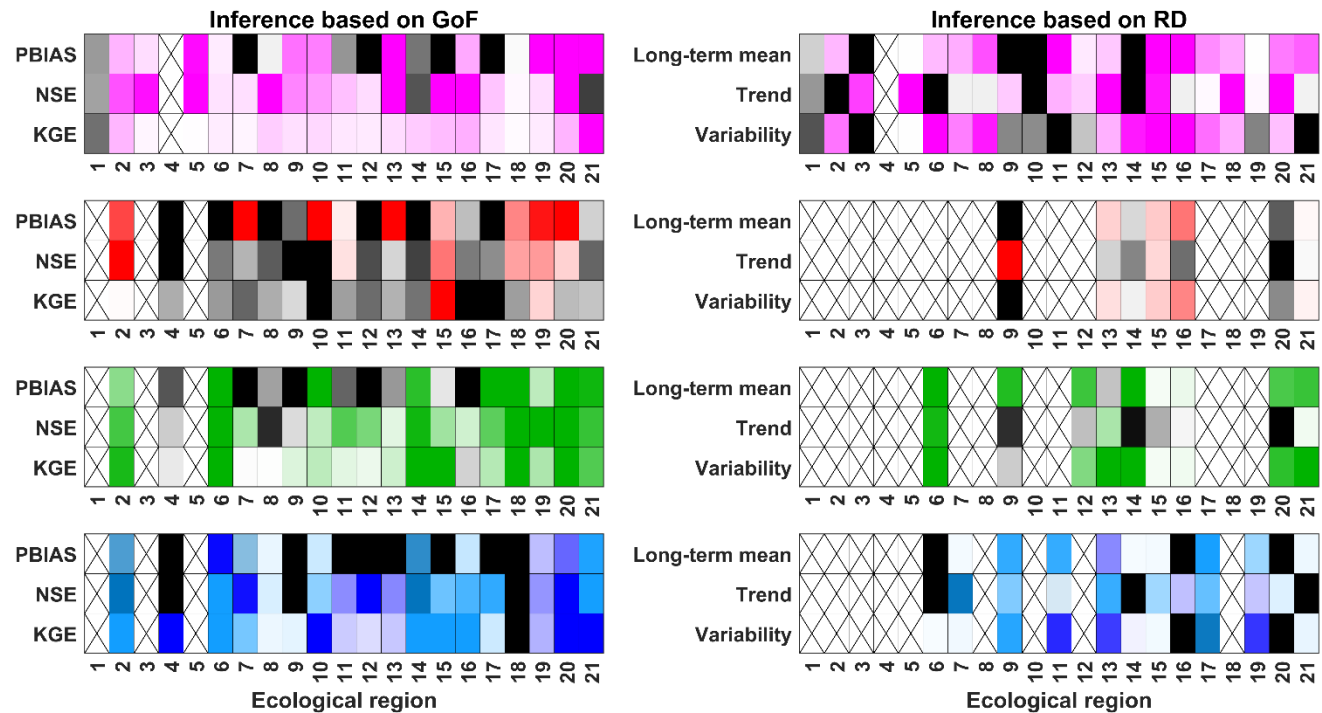


Figure 8: Relative fidelity of ERA5L and competing gridded representations based on the available observational constraints across ecological regions. The left and right columns show assessments based on the fidelity in GoFs for monthly time-series and RDs for annual characteristics, respectively. From top to bottom, rows correspond to SCE, Rho, SD, and SWE. Coloured cells indicate ecological regions where a competing gridded representation exhibits higher fidelity than ERA5L, whereas grey cells indicate regions where ERA5L exhibits higher fidelity than the competing representation(s). Colour saturation is proportional to the magnitude of the fidelity difference between ERA5L and the competing representation(s); white cells indicate regions where the available observational evidence does not distinguish their relative fidelity. Hatched cells denote ecological regions for which observational constraints are unavailable and relative fidelity therefore cannot be empirically assessed.

As noted above, diagnostic inconsistency refers to situations in which the inferred relative fidelity of competing gridded representations changes depending on the fidelity dimension and/or diagnostic metric considered. This issue becomes particularly important when the target characteristic cannot be evaluated directly because observational support is insufficient, leading to reliance on more readily calculable diagnostics as surrogates under the implicit assumption that higher fidelity according to those diagnostics also implies higher fidelity in the characteristic of interest. Figure 8 shows that this assumption



is not generally valid. Figure 9 illustrates the consequences of diagnostic inconsistency using three examples involving SCF, Rho, and SD. For SCF in the Northern Arctic, ECCC exhibits higher relative fidelity than ERA5L according to all three GoFs. In contrast, ERA5L exhibits higher fidelity when the comparison is based on RD in trend reconstruction. This reversal not only changes the preferred representation but also alters the inferred direction of change in SCF. A similar inconsistency occurs for Rho in the Taiga Shield, where ERA5L is preferred according to GoFs whereas CMC is preferred according to trend RD. The resulting distributions of gridded trends differ substantially between the two products, with CMC yielding a central tendency close to zero and considerably lower spatial variability. For SD in the Boreal Cordillera, CMC is preferred according to GoFs whereas ERA5L is preferred according to trend RD. Again, the resulting trend fields differ markedly, with CMC exhibiting substantially greater spatial variability and larger positive trends than ERA5L. Together, these examples demonstrate that diagnostic inconsistency can propagate directly into different representations of regional snow change. Where the characteristic of interest cannot itself be adequately constrained by observations, selecting a gridded product on the basis of time-series fidelity alone may therefore lead to a diverging inference from that obtained if fidelity can be evaluated directly for the target characteristic.

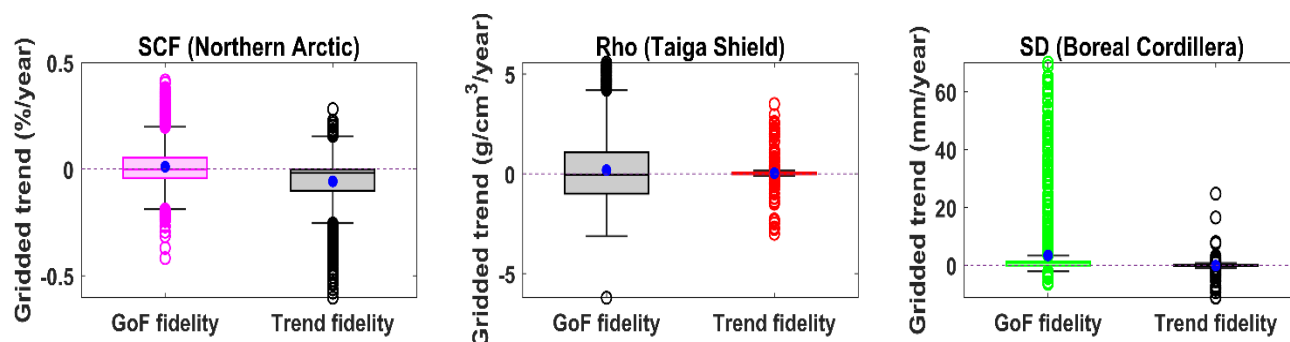


Figure 9: Differences in estimated gridded trends in annual mean SCF (left; Northern Arctic), Rho (middle; Taiga Shield), and SD (right; Boreal Cordillera) when the gridded representation is selected based on its relative fidelity according to GoFs versus trend reconstruction. Grey boxplots show trends inferred from ERA5L, whereas coloured boxplots show trends inferred from the corresponding competing representations (ECCC for SCF; CMC for Rho and SD). The blue dots in each boxplot show the inferred regional trend estimated using the mean of inferred gridded trends.

A second limitation arising from incomplete observational constraint is non-identifiability. Non-identifiability occurs when, for a given fidelity metric and target characteristic, the available observational evidence cannot meaningfully discriminate between two or more gridded representations, such that the evidence does not support preferring one representation over another. This limitation becomes consequential when representations that cannot be distinguished within observationally constrained portions of a region yield substantially different estimates over the larger observationally unconstrained domain.



Such cases occur in multiple product intercomparisons for Rho, SD, and SWE shown in Figure 8 and Figure S16 in the Supplement, under both GoF- and RD-based assessments. In such circumstances, apparent equivalence under benchmarking does not imply equivalence in the information ultimately inferred from the products. This condition is clearly illustrated for the annual trend in SD across the Marine West Coast Forest region (Figure 10). ERA5L and CMC cannot be meaningfully distinguished in their fidelity for reconstructing annual trend across the eight strongly conditioned grid cells. However, the two products diverge markedly across the observationally unconstrained grid cells, particularly in the northern part of the region, where ERA5L indicates predominantly moderate negative trends, whereas CMC indicates substantially stronger positive trends. Thus, two representations that cannot be discriminated where observations are available lead to conflicting inferences about the direction and magnitude of SD change where those observations are absent. Under such non-identifiability, selecting one representation while disregarding the other would imply a degree of certainty that is not supported by the available observational evidence.

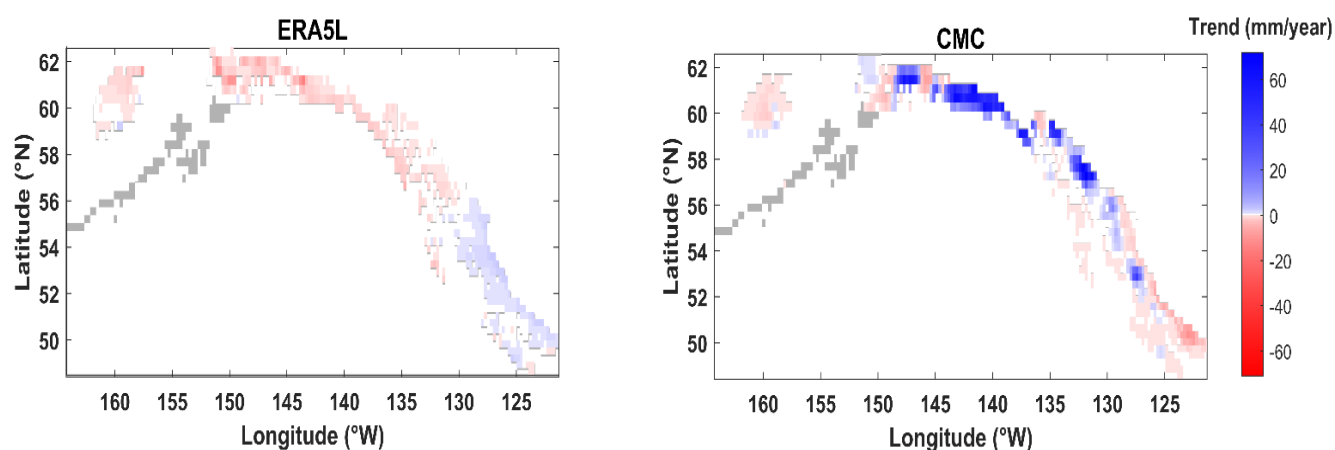
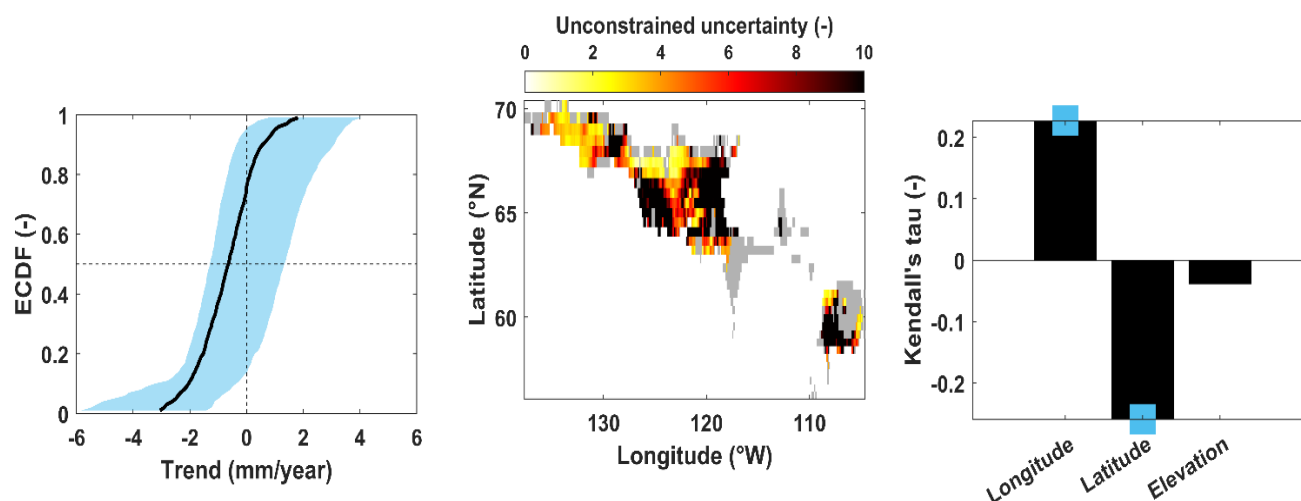


Figure 10: Non-identifiability of gridded trends in annual mean SD across the Marine West Coast Forest region. The left and right columns show trends inferred from ERA5L and CMC, respectively. Grey areas indicate grid cells for which trends could not be estimated due to missing data.

Unlike non-identifiability, where observational evidence exists but is insufficient to distinguish competing products, ambiguity arises where no observational constraint is available to evaluate their relative fidelity. This represents the strongest limit on empirical discrimination considered here because the competing representations cannot be ranked observationally within the region of interest. This situation affects all snow variables considered at both daily and monthly scales (Figures S17 and S18 in the Supplement). Under such circumstances, the available gridded representations should therefore be treated as unresolved alternative representations of snow characteristics. This condition is illustrated by trends in annual mean SWE across the



625 Southern Arctic, where the entire ecological region remains observationally unconstrained (Figure 11). ERA5L, CMC, ECCO, and AMSR therefore provide four unconstrained gridded representations of SWE trend across this approximately 800,000 km² region. The resulting pictures of regional SWE change differ substantially. Depending on the selected representation, between approximately 18% and 97% of grid cells exhibit negative SWE trends, with ERA5L indicating negative trends over about 80% of the region. Regional trend estimates also differ markedly in magnitude and even sign. The median gridded trend ranges
 630 from approximately -2 to 1.8 mm yr⁻¹ across the four products, whereas ERA5L indicates a median trend of nearly -1 mm yr⁻¹. Thus, the available products support fundamentally different interpretations of whether SWE is increasing or decreasing and by how much. The magnitude of this ambiguity, quantified using Eq. (5), can be up to one order of magnitude larger than the magnitude of the multi-product central trend estimate across parts of the central and southern Southern Arctic. Moreover, this uncertainty is itself spatially structured, increasing westward and decreasing northward, demonstrating that ambiguity is
 635 geographically heterogeneous. In practice, such ambiguity should not be resolved by selecting a single representation in the absence of specific observational evidence supporting that choice. As Figure 8 clearly shows, evidence of relative fidelity from one ecological region cannot be assumed transferable across ecological regions.



640 **Figure 11: Ambiguity in gridded trends in annual mean SWE across the Southern Arctic region.** Left, middle and right panels show spatial characteristics of the trends inferred from ERA5L, CMC, ECCO and AMSR, corresponding unconstrained uncertainty estimated from competing gridded representations, and spatial structure of unconstrained uncertainty, respectively. The solid black line in the left panel shows the ECDF of gridded trends inferred from ERA5L. Blue boxes overlaid on the bars in the right panel indicate statistically significant dependencies (p -value ≤ 0.05).



5 Discussion

645 5.1 Fidelity across snow variables and characteristics

Relative to the earlier analysis based primarily on comparison with a single gridded reference (Sarpong and Nazemi, 2025), the expanded multi-evidence benchmarking changes the scope of the conclusions while preserving several broad findings highlighted before. ERA5L generally shows higher fidelity for SCF and Rho than for SD and SWE, while CMC consistently exhibits higher daily and monthly time-series fidelity for SD. ERA5L SWE field also shows substantial monthly bias and
 650 relatively weak NSE and KGE performance compared to ECCO and CMC. These results are broadly consistent with previous evaluations reporting appreciable discrepancies in SD and SWE fields of ERA5 and ERA5L, particularly in complex terrain (Orsolini et al., 2019; Lei et al., 2023; Kouki et al., 2023; Monteiro and Morin, 2023; Scherrer et al., 2023), while the comparatively stronger Rho performance agrees with previous assessments (Gao et al., 2022). The earlier suggestion that disagreement in the ERA5L SWE field is more strongly associated with SD field than with Rho also remains qualitatively
 655 supported.

Characteristic fidelity, however, provides a less categorical picture. Our results clearly show that weak GoF fidelity does not preclude comparatively faithful reconstruction of selected long-term characteristics. ERA5L can perform poorly according to conventional GoFs yet remain competitive, or even provide the smallest discrepancy, for some annual characteristics. Most notably, ERA5L yields the smallest RD among the evaluated products for the annual long-term mean of SWE, particularly in
 660 mountainous grids, and can also show smaller trend discrepancies than competing representations for SD and SWE. ERA5L fidelity therefore depends on whether the application requires reproduction of the time series, climatological mean, interannual variability, or trend. This distinction helps reconcile apparently contrasting conclusions in the literature. Regional studies have identified substantial biases in ERA5L SD and SWE fields, whereas the recent SnowPEX+ assessment ranked ERA5L first overall among 23 Northern Hemisphere SWE products when climatology, variability, and trend were considered jointly
 665 (Mudryk et al., 2025). Our results show how both conclusions can be supported.

5.2 Dependence of fidelity on benchmarking diagnostics

The relationship between time-series and characteristic fidelity depends strongly on the characteristic considered. Long-term mean and detrended interannual variability generally exhibit the expected association with GoFs. Larger |PBIAS| tends to accompany larger |RD|, whereas higher NSE and KGE tend to accompany smaller |RD|. This is consistent with the sensitivity
 670 of these GoFs to mean behaviour, variability, and temporal correspondence (Gupta et al., 2009). Trend is markedly different. Trend is a comparatively small low-frequency component embedded within much larger day-to-day or month-to-month variability, so metrics characterizing correspondence across the full series are not necessarily informative about the trend component. Across both temporal scales, trend RD is not consistently associated with GoFs, and several cases reveal an explicit trade-off. For ERA5L daily SWE, for example, larger |PBIAS| is associated with smaller trend discrepancy, whereas higher



675 NSE and KGE are associated with larger trend RD. Thus, stronger apparent time-series fidelity can coincide with poorer representation of long-term change.

Divergence among fidelity dimensions or metrics constitutes diagnostic inconsistency, the first form of epistemic uncertainty considered here. Its consequences are evident in Figure 9, showing that representation favoured by GoFs differs from that favoured by trend RD, producing materially different trend distributions, and even reversal in the inferred direction of regional change. These findings extend the observation that no product performs best across all evaluation criteria (Mortimer et al., 2020) by showing that diagnostic choice can propagate directly into the physical interpretation obtained from a product. Where the target characteristic can be benchmarked directly, fidelity should therefore be evaluated for that characteristic. Where it cannot, disagreement among defensible diagnostic pathways should be retained as a diagnostic uncertainty rather than assuming that favourable GoF performance transfers to the target characteristic.

685 5.3 Epistemic uncertainty under incomplete observational constraints

A more fundamental limitation concerns whether available observations permit relative fidelity to be inferred at all. Despite combining MODIS with the extensive NorSWE archive, observational conditioning remains highly uneven across Canada and Alaska. This is consistent with long-recognized limitations in the spatial representativeness of in-situ snow measurements (Wang and Zheng, 2020), the terrain and scale dependence of gridded snow evaluation (Lei et al., 2023), and the heterogeneous footprints and temporal support of contemporary in-situ archives (Mortimer and Vionnet, 2025). The problem is most pronounced in northern regions, where gridded products are particularly valuable precisely because direct observations are scarce. Incomplete observations produce two additional limits on inference. Non-identifiability occurs when observations are available but do not meaningfully distinguish competing representations, even though those representations diverge elsewhere. The Marine West Coast Forest example shows why non-identifiability matters. In this example, ERA5L and CMC cannot be distinguished in their reconstruction of SD trend over the strongly conditioned grid cells, yet they infer markedly different and locally opposing trends over the larger unconstrained area. Apparent equivalence under benchmarking therefore does not imply equivalence in spatial inference outside the observational support.

Ambiguity arises when no observational constraint is available with which to discriminate the alternative representations. Across the Southern Arctic, four SWE products support markedly different trend patterns, magnitudes, and even signs, and their inter-product range can locally reach approximately one order of magnitude relative to the multi-product central signal. In this setting, product disagreement is not evidence that one product is wrong; it quantifies how unresolved the inferred characteristic remains. This is consistent with the broader epistemic argument that agreement or disagreement among model-based representations cannot establish the underlying natural state without independent evidence (Oreskes et al., 1994; Parker, 2016).



705 5.4 Implications for benchmarking and product selection

These epistemic limitations can themselves become informative outputs of benchmarking. Diagnostic inconsistency identifies characteristics for which conventional performance measures are inadequate surrogates. Application-specific diagnostics should therefore be preferred where observations permit; otherwise, the range produced by alternative defensible diagnostic choices can be retained as a diagnostic uncertainty in impact assessment and downstream studies.

710 Non-identifiability points instead to the need for additional discriminating evidence. This may include independent measurements, additional snow characteristics, cross-variable consistency among snow variables, or process-based constraints that challenge the products in ways not already represented by the existing metric. Such evidence is particularly relevant for snow because observational systems provide complementary rather than interchangeable information and constrain different aspects of the snow state (Hall and Riggs, 2007; Foster et al., 2005; Byun and Choi, 2014; Tsang et al., 2022). Until such
 715 evidence becomes available, non-identifiable representations should be retained jointly for impact assessment and downstream studies rather than forcing an unsupported ranking.

Ambiguity can similarly guide where new information is most valuable. The spread-to-signal measure identifies where alternative products disagree most strongly relative to the inferred signal. Regions with large ambiguity, especially where products disagree in sign, become candidates for enhanced monitoring; persistent disagreement in comparatively well-
 720 observed areas may instead point to greater value in improving model structure, forcing, assimilation, or retrieval procedures. Ambiguity can therefore help prioritize both monitoring and geographically targeted model improvement.

Taken together, these results suggest shifting the purpose of benchmarking away from identifying a universally “best” gridded snow product. More informative questions are which representation is supported for the variable and characteristic of interest given the observational constraints currently available, whether the available evidence can discriminate among alternatives,
 725 whether that conclusion is consistent across diagnostics, where unresolved uncertainty remains, and how this uncertainty should be handled and reduced. In this broader sense, benchmarking is better viewed as a process of evidential confirmation rather than verification. Its role is therefore to clarify the strength of support for competing representations and the limits of the inferences that can be drawn from them.

6 Conclusion and further remarks

730 This study reframes benchmarking of gridded snow products as an evidence-based inference problem rather than a comparison against an assumed reference truth. Using ERA5L as a test case, we harmonized SCF, Rho, SD, and SWE with competing gridded representations and observational constraints across 21 ecological regions of Canada and Alaska. The analysis



combined daily and monthly time-series diagnostics with relative differences in long-term mean, trend, and detrended interannual variability, while explicitly conditioning conclusions on the availability and strength of observational support.

735 Three main empirical conclusions emerge. First, ERA5L fidelity is strongly dependent on snow variable, temporal scale, target characteristic, and the observational and physiographic context of evaluation. ERA5L generally shows higher fidelity for SCF and Rho than for SD and SWE, and CMC consistently provides higher daily and monthly time-series fidelity for SD. ERA5L SWE also shows substantial monthly bias and weak NSE and KGE performance relative to several competing representations. Yet these time-series results do not translate into a universal ranking. ERA5L can remain competitive, or even provide the
740 smallest discrepancy, for selected long-term characteristics, including the annual long-term mean of SWE and trends of SD and SWE. Fidelity also varies with terrain, measurement technique, and observational density, although these effects are not uniform across variables or fidelity dimensions. Daily ERA5L SD and SWE time-series fidelity, for example, is generally lower in non-mountainous grids, whereas discrepancies in reconstructed long-term characteristics can be larger in mountainous grids, particularly for ERA5L SD. Daily SD and SWE performance also varies among measurement techniques, with lower
745 fidelity against mixed observations and snow pillows, while ERA5L SWE performs comparatively more favourably against passive-gamma observations at the monthly scale. Differences among products likewise vary across different levels of observational density. These results demonstrate that product fidelity cannot be characterized independently of the variable, characteristic, temporal scale, physiographic setting, and observational support through which it is evaluated.

Second, time-series fidelity and characteristic fidelity are not interchangeable. GoFs generally provide diagnostically
750 consistent information about the long-term mean and detrended interannual variability, but trend fidelity is weakly related to these diagnostics and can exhibit explicit trade-offs. A representation with stronger PBIAS, NSE, or KGE performance can therefore reconstruct trend less faithfully than a competing representation. This constitutes diagnostic inconsistency and is conceptually distinct from limitations in observational support, as it can arise even where observations fully constrain the comparison whenever different legitimate fidelity dimensions or diagnostic metrics lead to different rankings of competing
755 representations. In the examples examined here, diagnostic choice changes inferred regional trends and its variability. Therefore, where trend or another derived characteristic is the target quantity, fidelity should be evaluated directly for that characteristic whenever possible rather than inferred from generic time-series performance.

Third, observational support determines whether, and how strongly, competing representations can be discriminated. Where observations are available but do not meaningfully distinguish products that subsequently diverge outside the observational
760 support, their relative fidelity becomes non-identifiable. Where no observational constraint is available, ambiguity prevents empirical discrimination among the alternative representations altogether. These two limits are consequential rather than semantic. In the Marine West Coast Forest, ERA5L and CMC cannot be meaningfully distinguished in their reconstruction of SD trend over strongly conditioned grid cells, yet produce conflicting spatial patterns where observations are absent. In the Southern Arctic, the absence of observational constraint leaves four SWE representations supporting substantially different



765 trend magnitudes, spatial patterns, and direction. Such disagreement should therefore be retained as unresolved inferential uncertainty rather than eliminated through unsupported selection of a single representation.

Several limitations qualify the numerical results. MODIS and NorSWE data provide valuable observational constraints but are not error-free representations of the snow state and differ in spatial footprint, sampling frequency, record length, and representativeness relative to a 25 km grid. The competing products are also not fully independent. CMC incorporates in-situ
770 SD information, ECCC blends partly dependent model and retrieval products, and AMSR SWE depends on retrieval and snow-physics assumptions. The common 2002–2016 period, harmonization procedure, and pragmatic thresholds used for weak and strong conditioning can further influence the available support and estimated fidelity. These limitations mean that the specific rankings and uncertainty magnitudes reported here should not be transferred unconditionally beyond the evaluated domain, period, scale, and observational support.

775 The broader framework, however, is transferable. Diagnostic inconsistency can reveal sensitivity of an inferred characteristic to the choice of benchmarking criterion even where observational support is strong. In such conditions, disagreement among defensible diagnostics can be retained as diagnostic uncertainty rather than forcing a unique interpretation. Non-identifiability instead points to the need for additional discriminating evidence, including independent observations, complementary snow characteristics, cross-variable consistency, or process-based constraints. Ambiguity identifies regions where new observational
780 information is most valuable and can therefore help prioritize monitoring, while persistent disagreement in better-observed regions may point toward model, forcing, assimilation, or retrieval improvements. Future work should extend this approach to additional snow products, regions, and environmental variables and examine ways of propagating these distinct forms of inferential uncertainty into impact and attribution studies.

The central practical message is that gridded coverage should not be mistaken for evidential certainty, and closer agreement
785 with available observations cannot establish a representation as uniquely true (Oreskes et al., 1994). A useful benchmark should therefore state not only which representation performs better relative to other available representations, but also for which variable and characteristic, under what observational and diagnostic conditions that conclusion is supported, how consistent it is across diagnostics, and what remains unresolved, and how the uncertainty should be handled and reduced.

We hope this framework supports more informed and defensible use of ERA5L and other gridded snow products across
790 Canada, Alaska, and other cold regions globally.

Data availability

All data used in this study are publicly available. ERA5-Land data can be obtained at Climate Data Store, available at <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download>. CMC's SD/SWE product can be accessed from NASA's NSIDC at <https://nsidc.org/data/nsidc-0447/versions/1#anchor-data-access-tools>. ECCC SCF/SWE product can



795 be obtained from <https://climate-scenarios.canada.ca/?page=blended-snow-data>. AMSR-E/AMSR2 Unified Level-3 Global 5-Day 25 km SWE product is publicly available at https://nsidc.org/data/au_5dsno/versions/1. MODIS data can be also obtained from NASA's NSIDC at <https://nsidc.org/data/mod10cm/versions/61#anchor-documentation>. NorSWE data is available at <https://doi.org/10.5281/zenodo.15263370>. Elevation data can be obtained from USGS at <https://www.usgs.gov/coastal-changes-and-impacts/gmted2010>.

800 Supplement

There is Supplement file associated with this manuscript that is submitted along with the paper for peer review.

Competing interests

Authors do not have any competing interests.

Acknowledgments

805 Financial support of this research is provided by Canada's New Frontier Research Fund Exploration stream (NFRFE-2020-01298). To the memories of Mark Lanegan and Fariborz Vaziri, Iran's premier glaciologist.

References

- Akyurek, Z., Kuter, S., Karaman, Ç. H., and Akpınar, B.: Understanding the snow cover climatology over Turkey from ERA5-Land reanalysis data and MODIS snow cover frequency product, *Geosci.*, 13, 311, <https://doi.org/10.3390/geosciences13100311>, 2023.
- 810 Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W.: A suite of global, cross-scale topographic variables for environmental and biodiversity modeling, *Sci. Data*, 5, 180040, <https://doi.org/10.1038/sdata.2018.40>, 2018.
- Appel, I.: Uncertainty in satellite remote sensing of snow fraction for water resources management, *Front. Earth Sci.*, 12, 711–727, <https://doi.org/10.1007/s11707-018-0720-1>, 2018.
- 815 Backus, G. E. and Gilbert, J. F.: Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. Int.*, 13, 247–276, <https://doi.org/10.1111/j.1365-246X.1967.tb02159.x>, 1967.
- Bair, E. H., Rittger, K., Davis, R. E., Painter, T. H., and Dozier, J.: Validating reconstruction of snow water equivalent in California's Sierra Nevada using measurements from the NASA Airborne Snow Observatory, *Water Resour. Res.*, 52, 8437–8460, <https://doi.org/10.1002/2016WR018704>, 2016.
- 820 Balsamo, G., Viterbo, P., Beljaars, A., van den Hurk, B., Hirschi, M., Betts, A. K., and Scipal, K.: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System, *J. Hydrometeorol.*, 10, 623–643, <https://doi.org/10.1175/2008JHM1068.1>, 2009.



- 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.
- Baxter, R., Clelland, A., and Marshall, G.: Evaluating the performance of key ERA-Interim, ERA5, and ERA5-Land climate variables across Siberia, *Int. J. Climatol.*, 44, 2318–2342, <https://doi.org/10.1002/joc.8456>, 2024.
- Bitz, C. M., and Battisti, D. S.: Interannual to decadal variability in climate and the glacier mass balance in Washington, western Canada, and Alaska, *J. Climate*, 12, 3181–3196, [https://doi.org/10.1175/1520-0442\(1999\)012<3181:ITDVIC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<3181:ITDVIC>2.0.CO;2), 1999.
- Boelman, N. T., Liston, G. E., Gurarie, E., Meddens, A. J., Mahoney, P. J., Kirchner, P. B., ... and Vierling, L. A.: Integrating snow science and wildlife ecology in Arctic-boreal North America, *Environ. Res. Lett.*, 14, 010401, <https://doi.org/10.1088/1748-9326/aaec1>, 2019.
- Bormann, K. J., Westra, S., Evans, J. P., and McCabe, M. F.: Spatial and temporal variability in seasonal snow density, *J. Hydrol.*, 484, 63–73, <https://doi.org/10.1016/j.jhydrol.2013.01.032>, 2013.
- Brown, R. D. and Brasnett, B.: Canadian Meteorological Centre (CMC) Daily Snow Depth Analysis Data, Version 1, Boulder, Colorado USA, NASA National Snow and Ice Data Center Distributed Active Archive Center, <https://doi.org/10.5067/W9FOYWH0EQZ3>, 2010.
- Brown, R. D. and Frei, A.: Comment on “Evaluation of surface albedo and snow cover in AR4 coupled models” by A. Roesch, *J. Geophys. Res.-Atmos.*, 112, D22, <https://doi.org/10.1029/2006JD008339>, 2007.
- Brown, R. D., Brasnett, B., and Robinson, D.: Gridded North American monthly snow depth and snow water equivalent for GCM evaluation, *Atmos.-Ocean*, 41, 1–14, <https://doi.org/10.3137/ao.410101>, 2003.
- Brown, R. D., Smith, C., Derksen, C., and Mudryk, L.: Canadian in situ snow cover trends for 1955–2017 including an assessment of the impact of automation, *Atmos.-Ocean*, 59, 77–92, <https://doi.org/10.1080/07055900.2021.1911781>, 2021.
- Broxton, P. D. and van Leeuwen, W. J. D.: Structure from motion of multi-angle RPAS imagery complements larger-scale airborne lidar data for cost-effective snow monitoring in mountain forests, *Remote Sens.*, 12, 2311, <https://doi.org/10.3390/rs12142311>, 2020.
- Brubaker, K. L., Pinker, R. T., and Deviatova, E.: Evaluation and comparison of MODIS and IMS snow-cover estimates for the continental United States using station data, *J. Hydrometeorol.*, 6, 1002–1017, <https://doi.org/10.1175/JHM447.1>, 2005.
- Byun, K. and Choi, M.: Uncertainty of snow water equivalent retrieved from AMSR-E brightness temperature in northeast Asia, *Hydrol. Process.*, 28, 3173–3184, <https://doi.org/10.1002/hyp.9846>, 2014.
- Brun, E., Vionnet, V., Boone, A., Decharme, B., Peings, Y., Valette, R., Karbou, F., and Morin, S.: Simulation of Northern Eurasian local snow depth, mass, and density using a detailed snowpack model and meteorological reanalyses, *J. Hydrometeorol.*, 14, 203–219, <https://doi.org/10.1175/JHM-D-12-012.1>, 2013.
- Callaghan, T. V., Johansson, M., Brown, R. D., Groisman, P. Y., Labba, N., Radionov, V., et al.: Multiple effects of changes in Arctic snow cover, *Ambio*, 40, 32–45, <https://doi.org/10.1007/s13280-011-0213-x>, 2011.



- Danielson, J. J. and Gesch, D. B.: Global multi-resolution terrain elevation data 2010 (GMTED2010), US Geol. Surv. Open-File Rep., 2011-1073, <https://doi.org/10.3133/ofr20111073>, 2011.
- DeBeer, C. M., Wheeler, H. S., Carey, S. K., and Chun, K. P.: Recent climatic, cryospheric, and hydrological changes over the interior of western Canada: a review and synthesis, *Hydrol. Earth Syst. Sci.*, 20, 1573–1598, <https://doi.org/10.5194/hess-20-1573-2016>, 2016.
- Deems, J. S., Painter, T. H., and Finnegan, D. C.: Lidar measurement of snow depth: a review, *J. Glaciol.*, 59, 467–479, <https://doi.org/10.3189/2013JoG12J154>, 2013.
- Deng, H., Chen, Y., and Li, Y.: Glacier and snow variations and their impacts on regional water resources in mountains, *J. Geogr. Sci.*, 29, 84–100, <https://doi.org/10.1007/s11442-019-1585-2>, 2019.
- Di Marco, N., Avesani, D., Righetti, M., Zaramella, M., Majone, B., and Borga, M.: Reducing hydrological modelling uncertainty by using MODIS snow cover data and a topography-based distribution function snowmelt model, *J. Hydrol.*, 599, 126020, <https://doi.org/10.1016/j.jhydrol.2021.126020>, 2021.
- 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.
- Dong, C.: Remote sensing, hydrological modeling, and in situ observations in snow cover research: A review, *J. Hydrol.*, 561, 573–583, <https://doi.org/10.1016/j.jhydrol.2018.04.027>, 2018.
- Dong, J., Walker, J. P., and Houser, P. R.: Factors affecting remotely sensed snow water equivalent uncertainty, *Remote Sens. Environ.*, 97, 68–82, <https://doi.org/10.1016/j.rse.2005.04.010>, 2005.
- Du, J., Kimball, J. S., Jones, L. A., Kim, Y., Glassy, J., and Watts, J. D.: A global satellite environmental data record derived from AMSR-E and AMSR2 microwave Earth observations, *Earth Syst. Sci. Data*, 9, 791–808, doi:10.5194/essd-9-791-2017, 2017.
- Ermolaev, O. P., Mal'tsev, K. A., Mukharamova, S. S., Kharchenko, S. V., and Vedeneva, E. A.: Cartographic model of river basins of European Russia, *Geogr. Nat. Res.*, 38, 131–138, <https://link.springer.com/article/10.1134/S1875372817020032>, 2017.
- Fang, Y., and Leung, L. R.: Northern Hemisphere Snow Drought in Earth System Model Simulations and ERA5-Land Data in 1980–2014, *J. Geophys. Res.: Atmos.*, 128, e2023JD039308, <https://doi.org/10.1029/2023JD039308>, 2023.
- Fang, Y., Liu, Y., and Margulis, S. A.: A western United States snow reanalysis dataset over the Landsat era from water years 1985 to 2021, *Sci. Data*, 9, 677, doi:10.1038/s41597-022-01768-7, 2022.
- Fatolahzadeh Gheysari, A., Maghoul, P., Ojo, E. R., and Shalaby, A.: Reliability of ERA5 and ERA5-Land reanalysis data in the Canadian Prairies, *Theor. Appl. Climatol.*, 155, 3087–3098, <https://doi.org/10.1007/s00704-023-04771-z>, 2024.
- Foster, J. L., Sun, C., Walker, J. P., Kelly, R., Chang, A., Dong, J., and Powell, H.: Quantifying the uncertainty in passive microwave snow water equivalent observations, *Remote Sens. Environ.*, 94, 187–203, <https://doi.org/10.1016/j.rse.2004.09.012>, 2005.



- 890 Frei, A., Brown, R., Miller, J. A., and Robinson, D. A.: Snow mass over North America: Observations and results from the second phase of the atmospheric model intercomparison project, *J. Hydrometeorol.*, 6, 681–695, <https://doi.org/10.1175/JHM443.1>, 2005.
- Gao, S., Li, Z., Zhang, P., Zeng, J., Chen, Q., Zhao, C., et al.: An assessment of the applicability of three reanalysis snow density datasets over China using ground observations, *IEEE Geosci. Remote Sens. Lett.*, 19, 1–5, <https://doi.org/10.1109/LGRS.2022.3202897>, 2022.
- 895 Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. Climate*, 30, 5419–5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>, 2017.
- Gascoin, S., Barrou Dumont, Z., Deschamps-Berger, C., Marti, F., Salgues, G., López-Moreno, J. I., Revuelto, J., Michon, T., Schattan, P., and Hagolle, O.: Estimating fractional snow cover in open terrain from Sentinel-2 using the normalized difference snow index, *Remote Sens.*, 12, 2904, doi:10.3390/rs12182904, 2020.
- Gleason, K. E., and Nolin, A. W.: Charred forests accelerate snow albedo decay: Parameterizing the post-fire radiative forcing on snow for three years following fire, *Hydrol. Process.*, 30, 3855–3870, <https://doi.org/10.1002/hyp.10897>, 2016.
- 905 Global Modeling and Assimilation Office: MERRA-2 tavgM_2d_slv_Nx: 2d, monthly mean, time-averaged, single-level, assimilation, single-level diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), <https://doi.org/10.5067/AP1B0BA5PD2K>, 2016.
- Graham, R. M., Cohen, L., Ritzhaupt, N., Segger, B., Graversen, R. G., Rinke, A., Walden, V. P., Granskog, M. A., and Hudson, S. R.: Evaluation of six atmospheric reanalyses over Arctic sea ice from winter to early summer, *J. Clim.*, 32, 4121–4143, <https://doi.org/10.1175/JCLI-D-18-0643.1>, 2019.
- 910 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- 915 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.
- Hall, D. K., Foster, J. L., Verbyla, D. L., Klein, A. G., and Benson, C. S.: Assessment of snow-cover mapping accuracy in a variety of vegetation-cover densities in central Alaska, *Remote Sens. Environ.*, 66, 129–137, [https://doi.org/10.1016/S0034-4257\(98\)00051-0](https://doi.org/10.1016/S0034-4257(98)00051-0), 1998.
- 920 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](https://doi.org/10.1016/S0034-4257(02)00095-0), 2002.



- Hamm, A., Arndt, A., Kolbe, C., Wang, X., Thies, B., Boyko, O., et al.: Intercomparison of gridded precipitation datasets over a sub-region of the Central Himalaya and the Southwestern Tibetan Plateau, *Water*, 12, 3271, <https://doi.org/10.3390/w12113271>, 2020.
- 925 Hassler, B., and Lauer, A.: Comparison of reanalysis and observational precipitation datasets including ERA5 and WFDE5, *Atmosphere*, 12, 1462, <https://doi.org/10.3390/atmos12111462>, 2021.
- Hatami, S., and Nazemi, A.: A statistical framework for assessing temperature controls on landscape freeze-thaw: Application and implications in Québec, Canada (1979–2016), *J. Hydrol.*, 603, 126891, <https://doi.org/10.1016/j.jhydrol.2021.126891>, 2021.
- 930 Hatami, S., and Nazemi, A.: Compound changes in temperature and snow depth lead to asymmetric and nonlinear responses in landscape freeze-thaw, *Sci. Rep.*, 12, 2196, <https://doi.org/10.1038/s41598-022-06320-6>, 2022.
- Henn, B., Newman, A. J., Livneh, B., Daly, C., and Lundquist, J. D.: An assessment of differences in gridded precipitation datasets in complex terrain, *J. Hydrol.*, 556, 1205–1219, <https://doi.org/10.1016/j.jhydrol.2017.03.008>, 2018.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al.: The ERA5 global reanalysis, *Q. J. Roy. Meteor. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- 935 Huning, L. S., and AghaKouchak, A.: Global snow drought hot spots and characteristics, *Proc. Natl. Acad. Sci.*, 117, 19753–19759, <https://doi.org/10.1073/pnas.1915921117>, 2020.
- Hoppinen, Z., Palomaki, R. T., Brencher, G., Dunmire, D., Gagliano, E., Marziliano, A., Tarricone, J., and Marshall, H.-P.: Evaluating snow depth retrievals from Sentinel-1 volume scattering over NASA SnowEx sites, *The Cryosphere*, 18, 5407–5430, doi:10.5194/tc-18-5407-2024, 2024.
- 940 Kaemo, M., Hassanzadeh, E., and Nazemi, A.: A locally relevant framework for assessing the risk of sea level rise under changing temperature conditions: Application in New Caledonia, Pacific Ocean, *Sci. Total Environ.*, 834, 155326, <https://doi.org/10.1016/j.scitotenv.2022.155326>, 2022.
- Kazama, S., Izumi, H., Sarukkalgige, P. R., Nasu, T., and Sawamoto, M.: Estimating snow distribution over a large area and its application for water resources, *Hydrol. Process.*, 22, 2315–2324, <https://doi.org/10.1002/hyp.6826>, 2008.
- 945 Kelly, R. E., Chang, A. T., Tsang, L., and Foster, J. L.: A prototype AMSR-E global snow area and snow depth algorithm, *IEEE Trans. Geosci. Remote Sens.*, 41, 230–242, <https://doi.org/10.1109/TGRS.2003.809118>, 2003.
- Kendall, M. G.: A new measure of rank correlation, *Biometrika*, 30, 81–93, <https://doi.org/10.1093/biomet/30.1-2.81>, 1938.
- Kim, Y., Kimball, J. S., Du, J., Schaaf, C. L. B., and Kirchner, P. B.: Quantifying the effects of freeze-thaw transitions and snowpack melt on land surface albedo and energy exchange over Alaska and Western Canada, *Environ. Res. Lett.*, 13, 074028, <https://doi.org/10.1088/1748-9326/aacf72>, 2018.
- 950 Kim, Y., Kimball, J. S., Glassy, J., and Du, J.: An extended global Earth system data record on daily landscape freeze-thaw status determined from satellite passive microwave remote sensing, *Earth Syst. Sci. Data*, 9, 133–147, <https://doi.org/10.5194/essd-9-133-2017>, 2017.



- 955 Kouki, K., Luojus, K., and Riihelä, A.: Evaluation of snow cover properties in ERA5 and ERA5-Land with several satellite-based datasets in the Northern Hemisphere in spring 1982–2018, *The Cryosphere Discuss.*, 2023, 1–33, <https://doi.org/10.5194/tc-17-5007-2023>, 2023.
- Lackner, G., Domine, F., Nadeau, D. F., Parent, A.-C., Anctil, F., Lafaysse, M., and Dumont, M.: On the energy budget of a low-Arctic snowpack, *The Cryosphere*, 16, 127–142, <https://doi.org/10.5194/tc-2021-255>, 2022.
- 960 Lehning, M.: Snow–atmosphere interactions and hydrological consequences, *Adv. Water Resour.*, 55, 1–3, <http://dx.doi.org/10.1016/j.advwatres.2013.02.001>, 2013.
- Lei, Y., Pan, J., Xiong, C., Jiang, L., and Shi, J.: Snow depth and snow cover over the Tibetan Plateau observed from space in comparison with ERA5: Matters of scale, *Clim. Dyn.*, 60, 1523–1541, <https://doi.org/10.1007/s00382-022-06376-0>, 2023.
- Li, H., Wang, Z., He, G., and Man, W.: Estimating snow depth and snow water equivalence using repeat-pass interferometric SAR in the northern piedmont region of the Tianshan Mountains, *J. Sens.*, 2017, 8739598, <https://doi.org/10.1155/2017/8739598>, 2017.
- 965 Lindsay, R., Wensnahan, M., Schweiger, A., and Zhang, J.: Evaluation of seven different atmospheric reanalysis products in the Arctic, *J. Clim.*, 27, 2588–2606, <https://doi.org/10.1175/JCLI-D-13-00014.1>, 2014.
- Lievens, H., Demuzere, M., Marshall, H.-P., Reichle, R. H., Brucker, L., Brangers, I., de Rosnay, P., Dumont, M., Giroto, M., Immerzeel, W. W., Jonas, T., Kim, E. J., Koch, I., Marty, C., Saloranta, T., Schöber, J., and De Lannoy, G. J. M.: Snow depth variability in the Northern Hemisphere mountains observed from space, *Nat. Commun.*, 10, 4629, doi:10.1038/s41467-019-12566-y, 2019.
- Ma, H., Zhang, G., Mao, R., Su, B., Liu, W., and Shi, P.: Snow depth variability across the Qinghai Plateau and its influencing factors during 1980–2018, *Int. J. Climatol.*, 43, 1094–1111, <https://doi.org/10.1002/joc.7883>, 2023.
- 975 Maeda, T., Taniguchi, Y., and Imaoka, K.: GCOM-W1 AMSR2 Level 1R Product: Dataset of brightness temperature modified using the antenna pattern matching technique, *IEEE Trans. Geosci. Remote Sens.*, 54, 770–782, <https://doi.org/10.1109/TGRS.2015.2465170>, 2016.
- Marshall, H. P., and Koh, G.: FMCW radars for snow research, *Cold Reg. Sci. Technol.*, 52, 118–131, <https://doi.org/10.1016/j.coldregions.2007.04.008>, 2008.
- 980 Maurer, E. P., Rhoads, J. D., Dubayah, R. O., and Lettenmaier, D. P.: Evaluation of the snow-covered area data product from MODIS, *Hydrol. Process.*, 17, 59–71, <https://doi.org/10.1002/hyp.1193>, 2003.
- McMahon, G., Wiken, E. B., and Gauthier, D. A.: Toward a scientifically rigorous basis for developing mapped ecological regions, *Environ. Manage.*, 34, S111–S124, <https://doi.org/10.1007/s00267-004-0170-2>, 2004.
- Mitterer, C., and Schweizer, J.: Analysis of the snow-atmosphere energy balance during wet-snow instabilities and implications for avalanche prediction, *The Cryosphere*, 7, 205–216, <https://doi.org/10.5194/tc-7-205-2013>, 2013.
- 985 Molotch, N. P., Fassnacht, S. R., Bales, R. C., and Helfrich, S. R.: Estimating the distribution of snow water equivalent and snow extent beneath cloud cover in the Salt-Verde River basin, Arizona, *Hydrol. Process.*, 18, 1595–1611, <https://doi.org/10.1002/hyp.1408>, 2004.



- Monteiro, D., and Morin, S.: Multi-decadal analysis of past winter temperature, precipitation, and snow cover data in the European Alps from reanalyses, climate models, and observational datasets, *The Cryosphere*, 17, 3617–3660, <https://doi.org/10.5194/tc-17-3617-2023>, 2023.
- Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, *J. Geophys. Res. Atmos.*, 117, D08101, <https://doi.org/10.1029/2011JD017187>, 2012.
- Mortimer, C. and Mudryk, V.: Comments on Benchmarking Snow Fields of ERA5-Land in the Northern Regions of North America, <https://doi.org/10.5194/egusphere-2024-4150-CC1>, 2025.
- Mortimer, C., Mudryk, L., Derksen, C., Luoju, K., Brown, R., Kelly, R., and Tedesco, M.: Evaluation of long-term Northern Hemisphere snow water equivalent products, *The Cryosphere*, 14, 1579–1594, <https://doi.org/10.5194/tc-14-1579-2020>, 2020.
- Mortimer, C. and Vionnet, V.: Northern Hemisphere in situ snow water equivalent dataset (NorSWE, 1979–2021), *Earth Syst. Sci. Data*, 17, 3619–3640, doi:10.5194/essd-17-3619-2025, 2025.
- Mudryk, L. R., Derksen, C., Kushner, P. J., and Brown, R.: Characterization of Northern Hemisphere snow water equivalent datasets, 1981–2010, *J. Climate*, 28, 8037–8051, <https://doi.org/10.1175/JCLI-D-15-0229.1>, 2015.
- Mudryk, L., Mortimer, C., Derksen, C., Elias Chereque, A., and Kushner, P.: Benchmarking of snow water equivalent (SWE) products based on outcomes of the SnowPEX+ Intercomparison Project, *The Cryosphere*, 19, 201–218, <https://doi.org/10.5194/tc-19-201-2025>, 2025.
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., and Thépaut, J. N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, *Earth Syst. Sci. Data*, 13, 4349–4383, <https://doi.org/10.5194/essd-13-4349-2021>, 2021.
- Musselman, K. N., Addor, N., Vano, J. A., and Molotch, N. P.: Winter melt trends portend widespread declines in snow water resources, *Nat. Clim. Change*, 11, 418–424, <https://doi.org/10.1038/s41558-021-01014-9>, 2021.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.
- Nazemi, A., Wheeler, H. S., Chun, K. P., Bonsal, B., and Mekonnen, M.: Forms and drivers of annual streamflow variability in the headwaters of Canadian Prairies during the 20th century, *Hydrol. Process.*, 31, 221–239, <https://doi.org/10.1002/hyp.11036>, 2017.
- Niu, G. Y., and Yang, Z. L.: An observation-based formulation of snow cover fraction and its evaluation over large North American river basins, *J. Geophys. Res. Atmos.*, 112, D21106, <https://doi.org/10.1029/2007JD008674>, 2007.
- Notarnicola, C.: Hotspots of snow cover changes in global mountain regions over 2000–2018, *Remote Sens. Environ.*, 243, 111781, doi:10.1016/j.rse.2020.111781, 2020.
- Omernik, J. M., and Griffith, G. E.: Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework, *Environ. Manage.*, 54, 1249–1266, <https://doi.org/10.1007/s00267-014-0364-1>, 2014.



- Omernik, J. M.: Perspectives on the nature and definition of ecological regions, *Environ. Manage.*, 34, S27–S38, <https://doi.org/10.1007/s00267-003-5197-2>, 2004.
- Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, 263, 641–646, <https://doi.org/10.1126/science.263.5147.641>, 1994.
- Orsolini, Y., Wegmann, M., Dutra, E., Liu, B., Balsamo, G., Yang, K., and Arduini, G.: Evaluation of snow depth and snow cover over the Tibetan Plateau in global reanalyses using in situ and satellite remote sensing observations, *The Cryosphere*, 13, 2221–2239, <https://doi.org/10.5194/tc-13-2221-2019>, 2019.
- Pakoksung, K., and Takagi, M.: Effect of DEM sources on distributed hydrological model to results of runoff and inundation area, *Model. Earth Syst. Environ.*, 7, 1891–1905, <https://doi.org/10.1007/s40808-020-00914-7>, 2021.
- Parajka, J., and Blöschl, G.: The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models, *J. Hydrol.*, 358, 240–258, <https://doi.org/10.1016/j.jhydrol.2008.06.006>, 2008.
- Parker, W. S.: Reanalyses and observations: What’s the difference?, *Bull. Am. Meteorol. Soc.*, 97, 1565–1572, doi:10.1175/BAMS-D-14-00226.1, 2016.
- Patil, A., Singh, G., and Rüdiger, C.: Retrieval of snow depth and snow water equivalent using dual polarization SAR data, *Remote Sens.*, 12, 1183, doi:10.3390/rs12071183, 2020.
- Pulliainen, J.: Mapping of snow water equivalent and snow depth in boreal and sub-arctic zones by assimilating space-borne microwave radiometer data and ground-based observations, *Remote Sens. Environ.*, 101, 257–269, <https://doi.org/10.1016/j.rse.2006.01.002>, 2006.
- Rapaić, M., Brown, R., Markovic, M., and Chaumont, D.: An evaluation of temperature and precipitation surface-based and reanalysis datasets for the Canadian Arctic, 1950–2010, *Atmos. Ocean*, 53, 283–303, <https://doi.org/10.1080/07055900.2015.1045825>, 2015.
- Rasmussen, R. M., Chen, F., Liu, C. H., Ikeda, K., Prein, A., Kim, J., Schneider, T., Dai, A., Gochis, D., Dugger, A., Zhang, Y., Jaye, A., Dudhia, J., He, C., Harrold, M., Xue, L., Chen, S., Newman, A., Dougherty, E., Abolafia-Rosenzweig, R., Lybarger, N. D., Viger, R., Lesmes, D., Skalak, K., Brakebill, J., Cline, D., Dunne, K., Rasmussen, K., and Miguez-Macho, G.: CONUS404: The NCAR–USGS 4-km long-term regional hydroclimate reanalysis over the CONUS, *Bull. Am. Meteorol. Soc.*, 104, E1382–E1408, doi:10.1175/BAMS-D-21-0326.1, 2023.
- Riggs, G. A., Hall, D. K., and Román, M. O.: Overview of NASA’s MODIS and Visible Infrared Imaging Radiometer Suite (VIIRS) snow-cover Earth System Data Records, *Earth Syst. Sci. Data*, 9, 765–777, doi:10.5194/essd-9-765-2017, 2017.
- Rittger, K., Bormann, K. J., Bair, E. H., Dozier, J., and Painter, T. H.: Evaluation of VIIRS and MODIS snow cover fraction in High-Mountain Asia using Landsat 8 OLI, *Front. Remote Sens.*, 2, 647154, doi:10.3389/frsen.2021.647154, 2021.
- Saavedra, F. A., Kampf, S. K., Fassnacht, S. R., and Sibold, J. S.: Changes in Andes snow cover from MODIS data, 2000–2016, *The Cryosphere*, 12, 1027–1046, <https://doi.org/10.5194/tc-12-1027-2018>, 2018.
- Sarpong, R. and Nazemi, A.: Benchmarking Snow Fields of ERA5-Land in the Northern Regions of North America, *EGUsphere [preprint]*, <https://doi.org/10.5194/egusphere-2024-4150>, 2025.



- Sarpong, R., Nazemi, A., and AghaKouchak, A.: Creeping snow drought threatens Canada's water supply, *Commun. Earth Environ.*, 7, 143, doi:10.1038/s43247-025-03162-8, 2026.
- Scherrer, S. C., Göldi, M., Gubler, S., Steger, C. R., and Kotlarski, S.: Towards a spatial snow climatology for Switzerland: Comparison and validation of existing datasets, *Meteorol. Z.*, 33, 101–116, <https://doi.org/10.1127/metz/2023/1210>, 2023.
- Schilling, S., Dietz, A., and Kuenzer, C.: Snow Water Equivalent Monitoring—A Review of Large-Scale Remote Sensing Applications, *Remote Sens.*, 16, 1085, <https://doi.org/10.3390/rs16061085>, 2024.
- Schumacher, V., Justino, F., Fernández, A., Meseguer-Ruiz, O., Sarricolea, P., Comin, A., Peroni Venancio, L., and Althoff, D.: Comparison between observations and gridded data sets over complex terrain in the Chilean Andes: Precipitation and temperature, *Int. J. Climatol.*, 40, 5266–5288, <https://doi.org/10.1002/joc.6518>, 2020.
- Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, *J. Am. Stat. Assoc.*, 63, 1379–1389, <https://doi.org/10.1080/01621459.1968.10480934>, 1968.
- Sexstone, G. A., Clow, D. W., Stannard, D. I., and Fassnacht, S. R.: Comparison of methods for quantifying surface sublimation over seasonally snow-covered terrain, *Hydrol. Process.*, 30, 3373–3389, <https://doi.org/10.1002/hyp.10864>, 2016.
- Shijin, W., Yuande, Y., and Yanjun, C.: Global snow-and ice-related disaster risk: A review, *Nat. Hazards Rev.*, 23, 03122002, [https://doi.org/10.1061/\(ASCE\)NH.1527-6996.0000584](https://doi.org/10.1061/(ASCE)NH.1527-6996.0000584), 2022.
- Skaugen, T., Stranden, H. B., and Saloranta, T.: Trends in snow water equivalent in Norway (1931–2009), *Hydrol. Res.*, 43, 489–499, <https://doi.org/10.2166/nh.2012.109>, 2012.
- Stähli, M., Stacheder, M., Gustafsson, D., Schlaeger, S., Schneebeli, M., and Brandelik, A.: A new in situ sensor for large-scale snow-cover monitoring, *Ann. Glaciol.*, 38, 273–278, <https://doi.org/10.3189/172756404781814933>, 2004.
- Stiegler, C., Lund, M., Christensen, T. R., Mastepanov, M., and Lindroth, A.: Effects of interannual variability in snow accumulation on energy partitioning and surface energy exchange in a high-Arctic tundra ecosystem, *The Cryosphere Discuss.*, <https://doi.org/10.5194/tc-2016-5>, 2016.
- Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., and Lea, J.: Estimating snow water equivalent using snow depth data and climate classes, *J. Hydrometeorol.*, 11, 1380–1394, <https://doi.org/10.1175/2010JHM1202.1>, 2010.
- Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K. L.: A review of global precipitation data sets: Data sources, estimation, and intercomparisons, *Rev. Geophys.*, 56, 79–107, <https://doi.org/10.1002/2017RG000574>, 2018.
- Tedesco, M. and Jeyaratnam, J.: A new operational snow retrieval algorithm applied to historical AMSR-E brightness temperatures, *Remote Sens.*, 8, 1037, <https://doi.org/10.3390/rs8121037>, 2016.
- Tedesco, M. and Jeyaratnam, J.: AMSR-E/AMSR2 Unified L3 Global 5-Day 25 km EASE-Grid Snow Water Equivalent, Version 1, NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA, <https://doi.org/10.5067/0PX911G6417E>, 2019.
- Theil, H.: A rank-invariant method of linear and polynomial regression analysis, I, II, III, *Proc. K. Ned. Akad. Wet. Ser. A*, 53, 386–392, 521–525, 1397–1412, <https://dwc.knaw.nl/DL/publications/PU00018789.pdf>, 1950.



- 1090 Trepte, Q., Minnis, P., and Arduini, R. F.: Daytime and nighttime polar cloud and snow identification using MODIS data, in Optical Remote Sensing of the Atmosphere and Clouds III, vol. 4891, SPIE, 449–459, <https://doi.org/10.1117/12.467306>, 2003.
- Tsang, L., Durand, M., Derksen, C., Barros, A. P., Kang, D.-H., Lievens, H., Marshall, H.-P., Zhu, J., Johnson, J., King, J., Lemmetyinen, J., Sandells, M., Rutter, N., Siqueira, P., Nolin, A., Osmanoglu, B., Vuyovich, C., Kim, E., Taylor, D.,
- 1095 Merkouriadi, I., Brucker, L., Navari, M., Dumont, M., Kelly, R., Kim, R. S., Liao, T.-H., Borah, F., and Xu, X.: Review article: Global monitoring of snow water equivalent using high-frequency radar remote sensing, *The Cryosphere*, 16, 3531–3573, doi:10.5194/tc-16-3531-2022, 2022.
- Vionnet, V., Mortimer, C., Brady, M., Arnal, L., and Brown, R.: Canadian historical snow water equivalent dataset (CanSWE, 1928–2020), *Earth Syst. Sci. Data*, 13, 4603–4619, <https://doi.org/10.5194/essd-13-4603-2021>, 2021.
- 1100 Wang, Y., and Zheng, Z.: Spatial representativeness analysis for snow depth measurements of meteorological stations in northeast China, *J. Hydrometeorol.*, 21, 791–805, <https://doi.org/10.1175/JHM-D-19-0134.1>, 2020.
- Willeit, M., and Ganopolski, A.: The importance of snow albedo for ice sheet evolution over the last glacial cycle, *Clim. Past*, 14, 697–707, <https://doi.org/10.5194/cp-2017-122>, 2018.
- Wright, R. G., Murray, M. P., and Merrill, T.: Ecoregions as a level of ecological analysis, *Biol. Conserv.*, 86, 207–213,
- 1105 [https://doi.org/10.1016/S0006-3207\(98\)00002-0](https://doi.org/10.1016/S0006-3207(98)00002-0), 1998.
- Xu, J., Ma, Z., Yan, S., and Peng, J.: Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China, *J. Hydrol.*, 605, 127353, <https://doi.org/10.1016/j.jhydrol.2021.127353>, 2022.
- Yoon, Y., Kemp, E. M., Kumar, S. V., Wegiel, J. W., Vuyovich, C. M., and Peters-Lidard, C.: Development of a global
- 1110 operational snow analysis: The US Air Force Snow and Ice Analysis, *Remote Sens. Environ.*, 278, 113080, <https://doi.org/10.1016/j.rse.2022.113080>, 2022.
- You, J., Tarboton, D. G., and Luce, C. H.: Modeling the snow surface temperature with a one-layer energy balance snowmelt model, *Hydrol. Earth Syst. Sci.*, 18, 5061–5076, <https://doi.org/10.5194/hess-18-5061-2014>, 2014.