Propagating Meteorological Uncertainty in Physically Based Mountain Snow ~~Simulations~~Modeling

David R. Casson [1], Guoqiang Tang[2], Nicolás Vásquez[1], Andrew W. Wood[3,4], and Martyn P. Clark[1]

[1]Department of Civil Engineering, University of Calgary, Calgary, Canada
[2]State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan, China
[3]Climate and Global Dynamics, National Center for Atmospheric Research, Boulder, United States
[4]Civil and Environmental Engineering, Colorado School of Mines, Golden, United States

Correspondence: David R. Casson (dave.casson@ucalgary.ca)

Abstract.

Snow estimation in mountainous regions is uncertain because meteorology varies with topography, in situ observations are sparse and biased, and snow models ~~have~~ are affected by structural and scale limitations. Probabilistic methods are increasingly recognized as essential for quantifying and propagating these uncertainties in hydrological assessment. This study generates meteorological forcing ensembles designed for mountainous terrain and applies them for ~~probabilistic,~~ physically based snow modeling. Precipitation from global station datasets is corrected for ~~wind induced~~ wind-induced gauge undercatch using standardized transfer functions. Using ~~the Geospatial Probabilistic Estimation Package (GPEP), we generate station based meteorological ensembles that combine~~ geospatial probabilistic estimation, station observations are combined with static topographic predictors ~~with~~ and dynamic atmospheric predictors from reanalyses. Deterministic fields are estimated ~~with~~ using locally weighted regression and random forests, ~~and~~ while spatially correlated random fields ~~are used to sample residuals and~~ sample residuals to produce skillful, reliable ensembles. The resulting ensembles drive the SUMMA energy-balance snow model, calibrated using a Large Sample Emulator and applied in three contrasting basins: the Chena (Interior Alaska), Bow (Canadian Rockies), and Tuolumne (Sierra Nevada). Undercatch correction improves cold-season precipitation, dynamic predictors enhance regression skill, and random forests outperform locally weighted regression. Ensemble verification yields positive Brier Skill Scores, and ensemble-forced SUMMA simulations reproduce observed snow water equivalent with ~~credible~~ realistic magnitude, timing, and uncertainty~~estimates. This structured approach advances probabilistic estimation of mountain snow~~. This framework advances mountain snow estimation by explicitly representing forcing uncertainty in complex terrain, supporting physically based modeling, data assimilation, and water resource management~~applications~~.

1 Introduction

Measurements and model estimates of mountain snow are essential for predicting runoff, managing flood risk, and planning for water scarcity in snow-dependent regions (Huning and AghaKouchak, 2020; Immerzeel et al., 2020; Sturm et al., 2017; Wood and Lettenmaier, 2008). However, despite the recognized benefits and concerted research efforts, quantifying mountain snow continues to be a persistent and multi-faceted challenge characterized by considerable uncertainty (Girotto et al., 2020;

Largeron et al., 2020; Dozier et al., 2016). This challenge is driven by the influence of complex terrain on meteorology, snow distribution, and the inability to obtain observations that are spatially and temporally representative (Clark et al., 2011; Frei and Isotta, 2019). Snow models forced with meteorological data are also uncertain, as coarse-scale representations of both meteorology and snow dynamics fail to capture actual snowpack variability (Marsh et al., 2024). ~~Observations~~ The observations used to evaluate and constrain models are typically limited in spatial density and temporal frequency relative to the processes governing snow accumulation and melt (Sturm et al., 2010; Miller et al., 2022; Blöschl and Sivapalan, 1995). ~~The result is high~~ Consequently, uncertainty in mountain snow estimation ~~that often remains unquantified~~remains substantial and is rarely quantified explicitly.

Estimating meteorological variables in complex terrain has traditionally relied on interpolating sparse station observations into gridded fields while accounting for topographic effects. Early approaches focused on physiographic interpolation, using elevation and terrain descriptors to adjust station data for lapse rates and orographic gradients (Daly et al., 1994; Clark and Slater, 2006). Products such as the Parameter-elevation Regression on Independent Slopes Model (PRISM) follow this framework, combining station observations with expert-based weighting to represent climate fields in mountains (Daly et al., 2008, 2021). Complementary statistical approaches, including locally weighted regression, capture relationships between meteorological variables and static predictors such as elevation, slope, and aspect (Newman et al., 2020). ~~Gridded meteorological datasets are widely used in snow modeling, but can exhibit substantial bias and lack explicit uncertainty characterization, particularly in mountainous regions (Daly et al., 2021; Henn et al., 2018a). These biases propagate through snowpack simulations, making meteorological forcing a dominant source of model uncertainty (Raleigh et al., 2015). Global reanalyses such as ERA5 combine observations with numerical weather prediction to produce spatially and temporally complete datasets (Hersbach et al., 2020). While they~~ Global reanalyses offer consistency and coverage, they often underperform in complex terrain and require bias correction and downscaling before hydrological use (Beck et al., 2017; Hong et al., 2022; Berg et al., 2024; Lussana et al., 2018). Deterministic gridded products can overly smooth spatial variability and underestimate extremes (Ensor and Robeson, 2008; Gervais et al., 2014). Machine learning (ML) methods such as ~~Random Forests~~ random forests (RF) have also been successfully applied to resolve patterns of meteorological data in complex terrain (Wolfensberger et al., 2021; Baez-Villanueva et al., 2020). However, accurately resolving such variability requires far denser observational networks than currently exist (Frei and Isotta, 2019). Gridded meteorological datasets are widely used in snow modeling, but can exhibit substantial bias and lack explicit uncertainty characterization, particularly in mountainous regions (Daly et al., 2021; Henn et al., 2018a). These biases propagate through snowpack simulations, making meteorological forcing a dominant source of model uncertainty (Raleigh et al., 2015; Günther et al., 2019).

Snow ~~measurement is~~ observations are essential for understanding snowpack evolution and constraining models, but remains limited in spatial and temporal coverage. ~~In-situ~~ In situ networks, including snow pillows, automated sensors, and manual surveys, provide locally accurate estimates of snow depth and water equivalent, ~~yet are~~ but too sparse to capture heterogeneity driven by topography, wind, and vegetation (Pomeroy et al., 2007; Brown et al., 2019; Mortimer and Vionnet, 2025; Clark et al., 2011). Airborne and terrestrial lidar, along with satellite remote sensing, have expanded the ability to map high resolution snow depth, cover, and albedo across larger domains (Deems et al., 2013; Painter et al., 2016; Gascoin et al., 2024). However, their

use remains constrained by infrequent coverage, uncertainties in converting depth to water equivalent, and retrieval challenges in forested or complex terrain (Sturm et al., 2010; Margulis et al., 2019). Reanalysis and remote-sensing products complement ground-based observations by providing spatially continuous meteorological fields where ~~in-situ~~ in situ measurements are sparse or unreliable (Lundquist et al., 2019).

Snow models are widely used to quantify snow mass, codify process understanding and support hydrologic prediction (Pomeroy et al., 2007; Vionnet et al., 2021a; Essery et al., 2013; Sun et al., 2019; Marsh et al., 2024). Modeling in complex terrain presents challenges, including steep energy-balance gradients, heterogeneous snow processes, and systematically biased precipitation inputs (Zhao and Li, 2015; Clark et al., 2011). Approaches range from lumped conceptual degree-day models to physically based, high-resolution systems that ~~solve processes such as resolving energy~~ resolveenergy balance, blowing snow, canopy interception, and avalanching (Hrachowitz and Clark, 2017; Magnusson et al., 2015). However, increased model complexity does not necessarily improve predictive performance, ~~as~~ and meteorological forcing, structural and parameter uncertainties often dominate and compensate for unresolved processes (Menard et al., 2021). Model Intercomparison Projects (MIPs) demonstrate that inter-model variability arises not only from differing process representations but also from forcing data, calibration, and scale mismatches (Essery et al., 2013; Menard et al., 2021). Key processes such as refreezing, sublimation, wind redistribution, and rain-on-snow thermodynamics remain difficult to accurately simulate, yet strongly influence snow mass and energy budgets (Marks et al., 1998; Pomeroy et al., 2016; Barlage et al., 2010). As a result, snow model uncertainty reflects both incomplete process understanding and the difficulty in parameterizing subgrid variability in meteorological and terrain drivers (Günther et al., 2019). Increasingly, ensemble modeling, data assimilation, and high-resolution process studies are being used to constrain model states and improve predictive skill (Girotto et al., 2020; Largeron et al., 2020; Vionnet et al., 2021a; Sun and Margulis, 2025; Mott et al., 2023).

Ensemble approaches provide a means to represent and propagate uncertainty in both model inputs and model states. In meteorology and climate science, ensembles are used to characterize distributions of plausible outcomes (Collins, 2007), and hydrology has increasingly adopted ensemble methods to quantify predictive uncertainty and provide probabilistic streamflow forecasts (Cloke and Pappenberger, 2009; Wu et al., 2020). In mountain environments, where predictive skill depends critically on accurate representation of snowpack states, ensemble methods enable the integration of various observations through data assimilation to improve snow estimation (Clark et al., 2006; Wood et al., 2016; Vernay et al., 2015; Largeron et al., 2020; Girotto et al., 2020). Better characterization of precipitation and temperature uncertainty is particularly important, for example, given the increasing frequency of rain-on-snow events under a warming climate (Hotovy et al., 2025; Musselman et al., 2018). Recent developments in meteorological ensembles that incorporate both static topographic controls ~~to~~ on observations and dynamic atmospheric predictors offer promising pathways ~~to capture~~ for capturing input uncertainty (Clark and Slater, 2006; Newman et al., 2020; Bunn et al., 2022; Tang et al., 2024). This evolution from deterministic interpolation to probabilistic ensemble generation reflects a growing recognition that uncertainty quantification is essential in mountainous terrain (Frei and Isotta, 2019; Bertoncini and Pomeroy, 2025; Kim et al., 2021; Mott et al., 2023).

A key gap remains in how probabilistic meteorological information is generated and evaluated for mountain snow modeling. Most existing studies in complex terrain focus on developing or validating probabilistic precipitation and temperature estimates

(e.g., Frei and Isotta, 2019; Bertoncini and Pomeroy, 2025), yet few extend this uncertainty representation through to physically based snow modeling. Consequently, the influence of forcing uncertainty on simulated snowpack evolution, melt timing, and water availability remains poorly quantified (Dozier et al., 2016; Largeron et al., 2020; Girotto et al., 2020; Tang et al., 2025a). Mountain snow modeling studies using physically based or energy-balance approaches very often continue to rely on a single deterministic meteorological forcing, typically derived from gridded observation-based or downscaled reanalysis products (Essery et al., 2013; Vionnet et al., 2021a; Berg et al., 2024; Marsh et al., 2024). Many studies evaluating snow ensembles are using "ensembles of opportunity" assembled from independent reanalysis datasets, which capture the diversity of datasets but not the true uncertainty (Schreiner-McGraw and Ajami, 2022; Henn et al., 2018b; Kim et al., 2021; Tyson et al., 2023; Casson et al., 2018).

This study directly addresses that gap by linking station-based probabilistic meteorology with physically based snow modeling. The workflow explicitly corrects for precipitation undercatch, generates spatially and temporally coherent ~~high resolution~~ high-resolution meteorological ensembles, and propagates these uncertainties through the SUMMA energy-balance snow model. By testing this approach across three contrasting mountain basins, the Chena (Interior Alaska), Bow (Canadian Rockies), and Tuolumne (Sierra Nevada), ~~and~~ using large domain datasets, this study demonstrates a transferable approach for representing and evaluating forcing uncertainty through to physically based snow modeling in complex terrain. The remainder of this paper is structured as follows. Section 2 describes the study areas and river basins. Section 3 describes the datasets used to construct and evaluate the ensembles and snow simulations. Section 4 describes the ensemble forcing methodology, including undercatch correction, predictor screening, and ensemble generation using the Geospatial Probabilistic Estimation Package (GPEP; (Tang et al., 2024)), followed by the process-based modeling approach. Section 5 presents the evaluation of the ensemble in the three basins, Section 6 discusses the implications of constructing mountain meteorological ensembles and snow modeling, and Section 7 summarizes key findings and future directions.

2 Study Area

Three river basins were selected to represent a broad range of hydroclimatic, physiographic, and monitoring contexts: the Chena above Two Rivers (Alaska), the Bow above Calgary (Alberta), and the Tuolumne above Hetch Hetchy (California) (hereafter Chena, Bow, and Tuolumne; Figure 1). Snow observation datasets in the basins have comparable station densities~~,~~; however, they differ in meteorological observational networks, climate regime, and hydrological landscape. These contrasts encompass subarctic, continental mountain, and ~~maritime influenced~~ maritime-influenced high-elevation environments. Together, these basins provide a set of testbeds that capture the diversity of conditions under which ensemble meteorological forcing systems for mountainous snow modeling must operate.

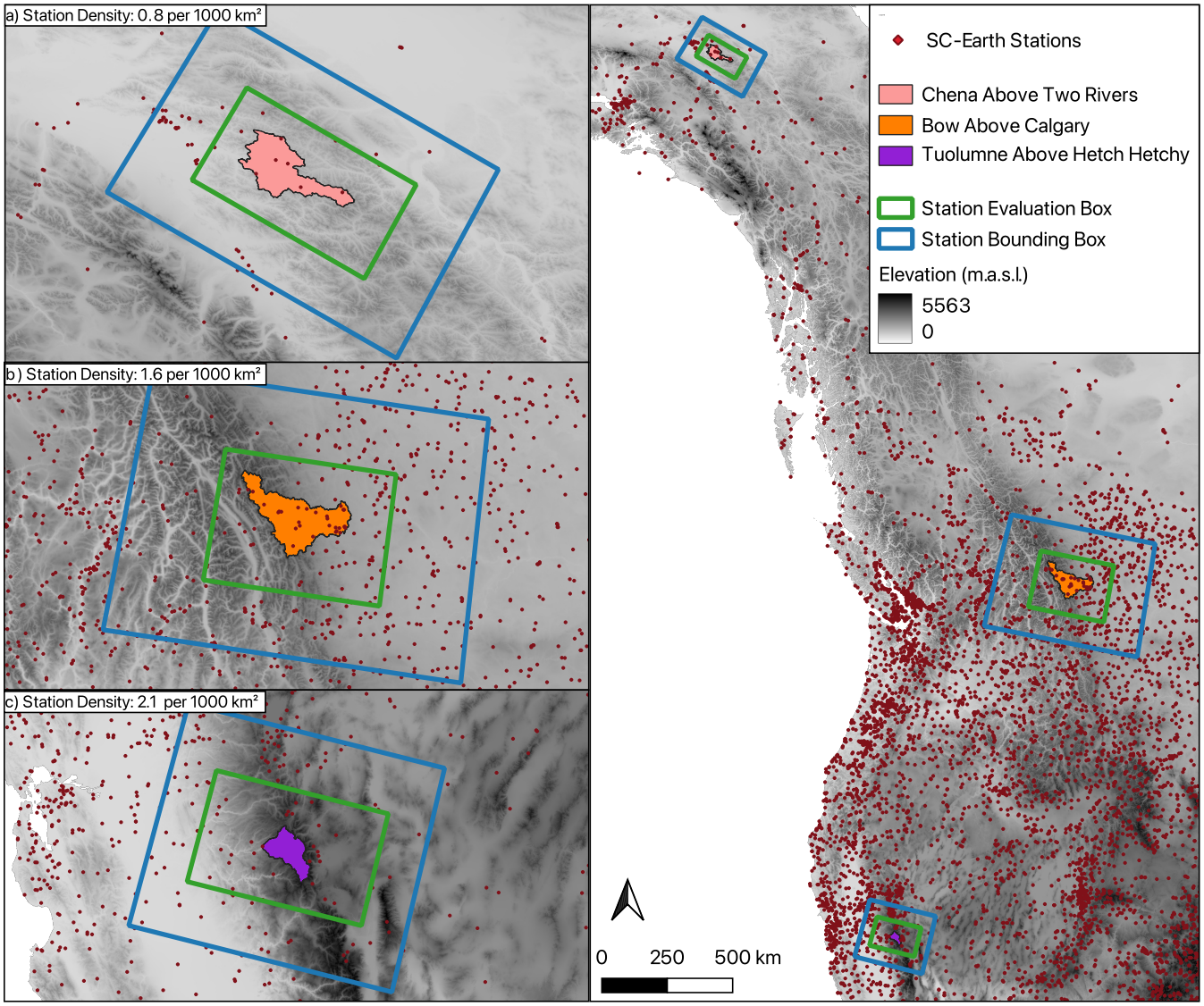


Figure 1. Chena, Bow, and Tuolumne river basins shown with SC-Earth meteorological stations. The larger station bounding box (blue) delineates the stations used to build ensembles and the area for grid generation. The smaller station evaluation box (green) delineates the stations used to evaluate meteorological ensemble performance. The elevation is shown as meters above sea level (m.a.s.l.) from MERIT Hydro.

The Chena is a tributary of the Tanana River in Interior Alaska, flowing westward towards Fairbanks. The Chena originates in the Yukon-Tanana Uplands, a region of rolling hills and mountains, with vegetation characterized by a mixture of boreal
125 forest and permafrost (Frey et al., 1970; Anderson, 1970). The basin covers an area of approximately 5,500 km^2, with an elevation ranging from approximately 100 m.a.s.l. at the outlet to 1,600 m.a.s.l. at its highest point (Vuyovich and Daly, 2012). It contains discontinuous permafrost, with the highest concentrations occurring at higher elevations. The watershed experiences

a subarctic climate, characterized by cold, dry winters and warm, relatively humid summers. Annual precipitation averages 380 to 510 mm per year ~~with its hydrology that~~ and basin hydrology is dominated by snowmelt runoff, though the largest recorded floods have occurred in summer due to intense rainfall (Liao and Zhuang, 2017). Snow observations in the Chena region are challenging due to difficulties in accurately measuring solid precipitation and the sparse distribution of observation stations across a large terrain ~~Bennett et al. (2019)~~(Bennett et al., 2019). Due to previous flood events, infrastructure developments and location above Fairbanks, there are dedicated snow monitoring stations and programs in the Chena. The majority of snow monitoring stations are part of the Natural Resources Conservation Service (NRCS) Snow Telemetry (SNOTEL) program (Fleming et al., 2023). The NRCS also conducts snow surveys twice per year at the start of April and May at several locations throughout the basin (Vuyovich and Daly, 2012).

The Bow is located in the Eastern Canadian Rocky Mountains in Southern Alberta. It flows south-east through Banff National Park and the foothills, then meanders eastward through the Prairies towards its confluence with the Oldman River (Tanzeeba and Gan, 2012). The Bow basin covers approximately 25,000 km^2 and begins at Bow Lake at 1,920 m.a.s.l., then drops to 1,180 m.a.s.l. at its confluence with the Oldman River (BRCB, 2010). The Bow has a semi-arid climate with precipitation of roughly 400-500 mm per year. Snowmelt runoff accounts for roughly 80% of the Bow River inflow, the rest coming from rainfall, groundwater, and a small (3%) contribution from glacier melt (Bash and Marshall, 2014). The mountainous regions of the watershed are a dominant source of surface water, and exert control over flood-to-drought conditions (Vionnet et al., 2020). Snow monitoring in the Bow ~~is a combination of~~ combines automated sites and manual snow surveys focused on late winter~~, May, and April~~. The measurements have been consolidated in the Canadian historical SWE dataset (CanSWE) which provides a detailed QA/QC dataset across Canada (Vionnet et al., 2021b). Bow watershed meteorological stations in the Environment and Climate Change Canada (ECCC) surface network have undergone undercatch evaluation as part of a modernization and shift from manual to automatic stations (Smith et al., 2022).

The Tuolumne is located in the Sierra Nevada in California and flows from high-elevation headwaters in Yosemite National Park to its confluence with the San Joaquin River. The Tuolumne spans 700–3,900 m.a.s.l. in elevation, covering approximately 1,650 km^2. Tuolumne streamflow is primarily driven by seasonal snowmelt, supporting downstream ecosystems, agriculture, and urban water supplies. Approximately 80–90% of the annual precipitation (800–~~1000~~ 1,000 mm per year) occurs as snow during winter months, though both precipitation amount and timing show considerable seasonal and year-to-year variability (Cristea et al., 2014). The Tuolumne Above Hetch Hetchy has low human intervention, such as dams or diversions, and land-use changes (Rice et al., 2011). The snow distribution in the Tuolumne River Basin is relatively similar from year to year, mostly controlled by strong topography and similar storm tracks controlling snow deposition patterns (Pflug and Lundquist, 2020; Lundquist et al., 2015a). The Tuolumne has been extensively studied for snow distribution due to its importance for water supply and the availability of high-resolution snow-depth measurements from airborne lidar (Painter et al., 2016). The combination of scanning lidar, an imaging spectrometer, and physically based modeling has produced a unique dataset of high-resolution snow water equivalent and snow albedo maps. ~~This~~ These spatially detailed data ~~has~~ have led to numerous studies assessing snow quantity and distribution (Margulis et al., 2019; Currier and Lundquist, 2018; Pflug and Lundquist, 2020), as well as water balance and runoff estimation (Lundquist et al., 2015b; Henn et al., 2016, 2018c; Hedrick et al., 2020).

3 Data

All datasets used in this study were selected and processed in accordance with FAIR data principles (Wilkinson et al., 2016). The analysis integrates publicly available meteorological, topographic, and snow observation datasets, each documented with open metadata and standardized formats to promote transparency and interoperability.

3.1 Meteorological Station Data

This study uses a Serially Complete (SC) station dataset that includes daily precipitation, ~~temperature mean~~ mean temperature, temperature range and wind speed with global spatial coverage from 1950-2019 (Tang et al., 2021). This SC-Earth station dataset is derived primarily from raw data from the Global Historical Climatology Network–Daily (GHCN-D; Durre et al., 2010) and the Global Surface Summary of the Day (GSOD; Menne et al., 2012). ERA5 reanalysis data are also used in SC-Earth production to correct a timeshift issue and assist with global gap filling. GHCN-D has been widely used for the production, evaluation, and correction of gridded meteorological datasets (Contractor et al., 2020; Beck et al., 2017; Hong et al., 2022; Tang et al., 2020). It includes precipitation and temperature data that undergo a comprehensive, automated quality assurance process (Durre et al., 2010). GSOD provides complementary wind speed observations used here to implement a precipitation undercatch correction strategy, though its quality control protocols are less rigorous, and caution is advised when using GSOD and SC-Earth wind estimates (Tang et al., 2021). GHCN-D, GSOD, and SC-Earth datasets do not include undercatch adjustments.

Station data compiled in SC-Earth originate from diverse networks employing various gauge configurations. In the Bow watershed, for instance, multiple monitoring networks use different gauge types even within the same region (Bertoncini and Pomeroy, 2025). Across the Chena, Bow, and Tuolumne basins, meteorological observations come from numerous providers lacking detailed gauge metadata. Although additional research datasets exist for these basins, they were not examined in this study due to limited temporal coverage, ~~study specific~~ study-specific assumptions and lack of serial completeness (Daly and Vuyovich, 2012; Fang et al., 2019; Lundquist et al., 2016; Henn et al., 2016; Bertoncini and Pomeroy, 2025).

3.2 Topographic and Land Surface Data

The MERIT Digital Elevation Model (DEM) is used for multiple applications in this study, including the derivation of topographic descriptors for geospatial estimation. MERIT DEM is available globally at 3 arcsecond (90-m at the equator) horizontal resolution (Yamazaki et al., 2017). In a previous study using the GPEP methodology, descriptors such as elevation, latitude, longitude, and west-east and south-north slopes (which proxy for terrain aspect) were utilized (Newman et al., 2015). This current study assesses these predictors, and also examines aspect, slope, and topographic position index (TPI). ~~Smoothing of topographic predictors with a Gaussian 2D~~ Topographic predictors were also smoothed with a two dimensional Gaussian filter over different radii ~~was also performed~~ to evaluate the relative skill as a regression predictor.

TanDEM-X Hydro (TDX-Hydro), a catchment and vector stream network dataset, was used for the spatial discretization ~~for snow modeling~~of the hydrological model, as shown in Figure 2. This dataset was derived from a high-resolution (12m)

TanDEM-X elevation data conditioned with vegetation bias removal, adaptive smoothing, river enforcement, and depression handling, resulting in a highly detailed representation of global hydrography (Carlson et al., 2024). The underlying ~~TanDEM-x~~ TanDEM-X elevation data are not publicly available, however, TDX-Hydro discretizations are. The use of a ~~high resolution~~ high-resolution catchment dataset helps capture topographic details such as elevation more accurately. Working with smaller catchments, the dataset can capture terrain variations that coarser datasets might miss; however, there is a trade off between spatial resolution and computational time in modeling (Chaney et al., 2018). For consistency between river basins, the discretization approach taken does not explicitly represent some topographic controls on snow, such as slope and aspect.

Multiple geospatial datasets are used to support hydrological model parameterization and the calculation of catchment attributes~~for modeling such as~~ , including land cover and aspect. Land cover characteristics are obtained from the MODIS MCD12Q1 dataset (Friedl and Sulla-Menashe, 2019), and ~~the~~ soil properties are sourced from the USDA NRCS database (Knoben, 2021) derived from the SoilGrids 250 m dataset (Hengl et al., 2017). Further details on these datasets applied for hydrological modeling are available in Knoben et al. (2022).

3.3 Snow ~~Observation~~ Data

3.4 In Situ Snow Data

The Northern Hemisphere Historical In Situ Snow Water Equivalent Dataset (NorSWE) provides a comprehensive record of snow water equivalent (SWE) observations spanning 1979–2021 (Mortimer and Vionnet, 2025). The NorSWE dataset integrates data from nine sources across North America, Russia, and Finland. It includes SWE measurements from manual snow surveys, snow pillows, airborne and ground-based passive gamma radiation sensors, with snow depth (SD) data included when available. Bulk snow density is derived where both SWE and SD are present. Quality control procedures, adapted from (Vionnet et al., 2021b), include variable-specific range checks and elevation verification using the USGS National Elevation Dataset where applicable. ~~Although additional snow monitoring products, including airborne lidar, are available for the Tuolumne and to a lesser extent the Bow, the analysis is constrained to in-situ observations for consistency.~~

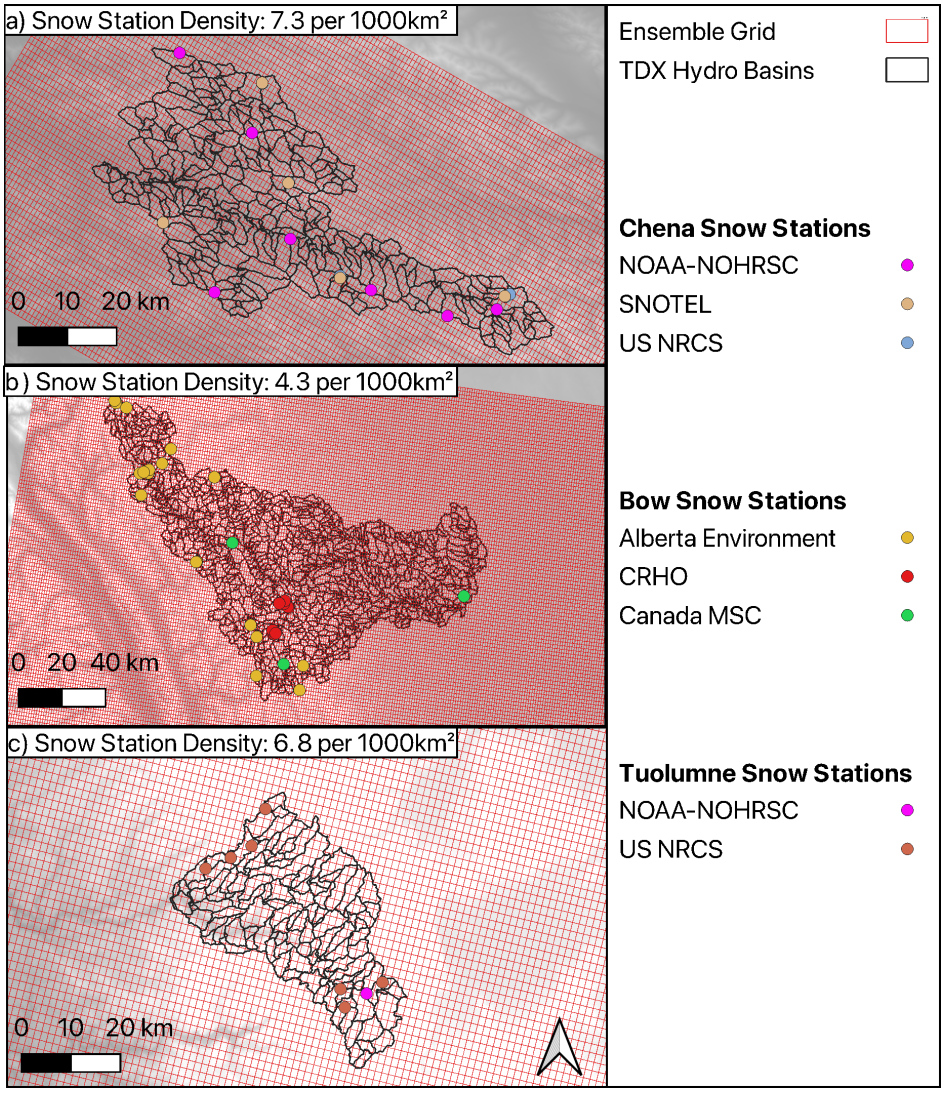


Figure 2. Chena, Bow, and Tuolumne River Basins showing the meteorological forcing grid (1/64°) in red, the TDX Hydro basin delineations in black, and the snow stations from the multi-source NorSWE dataset. Data sources include National Oceanic and Atmospheric Administration - National Operational Hydrologic Remote Sensing Center (NOAA-NOHRSC), SNOwpack TELemetry Network (SNOTEL), United States Natural Resources Conservation Service (US NRCS), Canadian Rockies Hydrological Observatory (CRHO) and Canada Meterological Services of Canada (Canada MSC).The NRCS datasource compile observations from state level collection offices, including California Department of Water Resources.

For the calibration of snow parameters using a Large Sample Emulator (LSE) approach, station data from Mortimer and Vionnet (2025) and ~~(Sun et al., 2019)~~ Sun et al. (2019) are used. For analysis of the undercatch correction procedures, the ~~quality controlled~~ quality-controlled dataset of Smith et al. (2022) was used for comparative purposes.

3.5 Remotely Sensed and Lidar Snow Data

Spatial snow cover was characterized using the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra Snow Cover Daily L3 Global 500 m product (MOD10A1, Collection 6.1), which provides daily global snow cover observations derived from the Normalized Difference Snow Index (NDSI) (Hall and Riggs, 2021). The MODIS product was used to derive spatial metrics that describe the extent and seasonal evolution of snow cover.

Spatial snow water equivalent (SWE) observations were obtained from the Airborne Snow Observatory (ASO) L4 Lidar Snow Water Equivalent 50 m product for the Tuolumne (Painter, 2018). The ASO program combines airborne lidar measurements of snow depth with physically based snow density modeling to produce high-resolution SWE estimates across mountainous watersheds (Painter et al., 2016). These data provide spatially distributed observations for evaluating simulated SWE patterns.

3.6 Meteorological Reanalyses

The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) is a ~~state-of-the-art~~ state of the art global reanalysis dataset that provides consistent and high-quality estimates of atmospheric, land, and oceanic variables from 1950 to the present (Hersbach et al., 2020). In this study, we use ERA5 atmospheric (meteorological) forcing variables from the full reanalysis product, rather than the higher-resolution ERA5-Land dataset. ERA5 combines a vast array of historical observations with advanced numerical weather prediction models through data assimilation techniques to produce a spatial resolution of 0.25 degrees and hourly temporal resolution. In this study, we use wind and temperature from ERA5 for station precipitation undercatch correction when co-located measurement data is not available. Additionally, we use ERA5 as a benchmark forcing dataset for evaluating snow modeling performance within the Structure for Unifying Multiple Modeling Alternatives (SUMMA) hydrological modeling framework. (Clark et al., 2015a, c, b).

The Canadian Surface Reanalysis (CaSR) is a high-resolution, observation-constrained reanalysis dataset developed to provide detailed surface meteorological fields over North America (Gasset et al., 2021). The most recent version, CaSR v3.1, incorporates enhancements to the data assimilation system, expanded observational networks, and updated model physics to improve the representation of near-surface atmospheric conditions. CaSR provides meteorological variables at a 10 km spatial resolution and hourly temporal frequency, making it well-suited for regional hydrometeorological studies. In this study, CaSR is used as a source of dynamic predictor data for geospatial estimation and as a benchmark dataset for evaluating snow modeling performance within the SUMMA.

4 Methods

Reproducibility has historically been a challenge in hydrology and water resources research (Stagge et al., 2019). Recent efforts have emphasized open, transparent, and automated workflows in the Earth sciences (e.g., Gil et al., 2016; Knoben et al., 2022; Arnal et al., 2024). This study follows a structured workflow, which is illustrated at a high level in Figure 3.

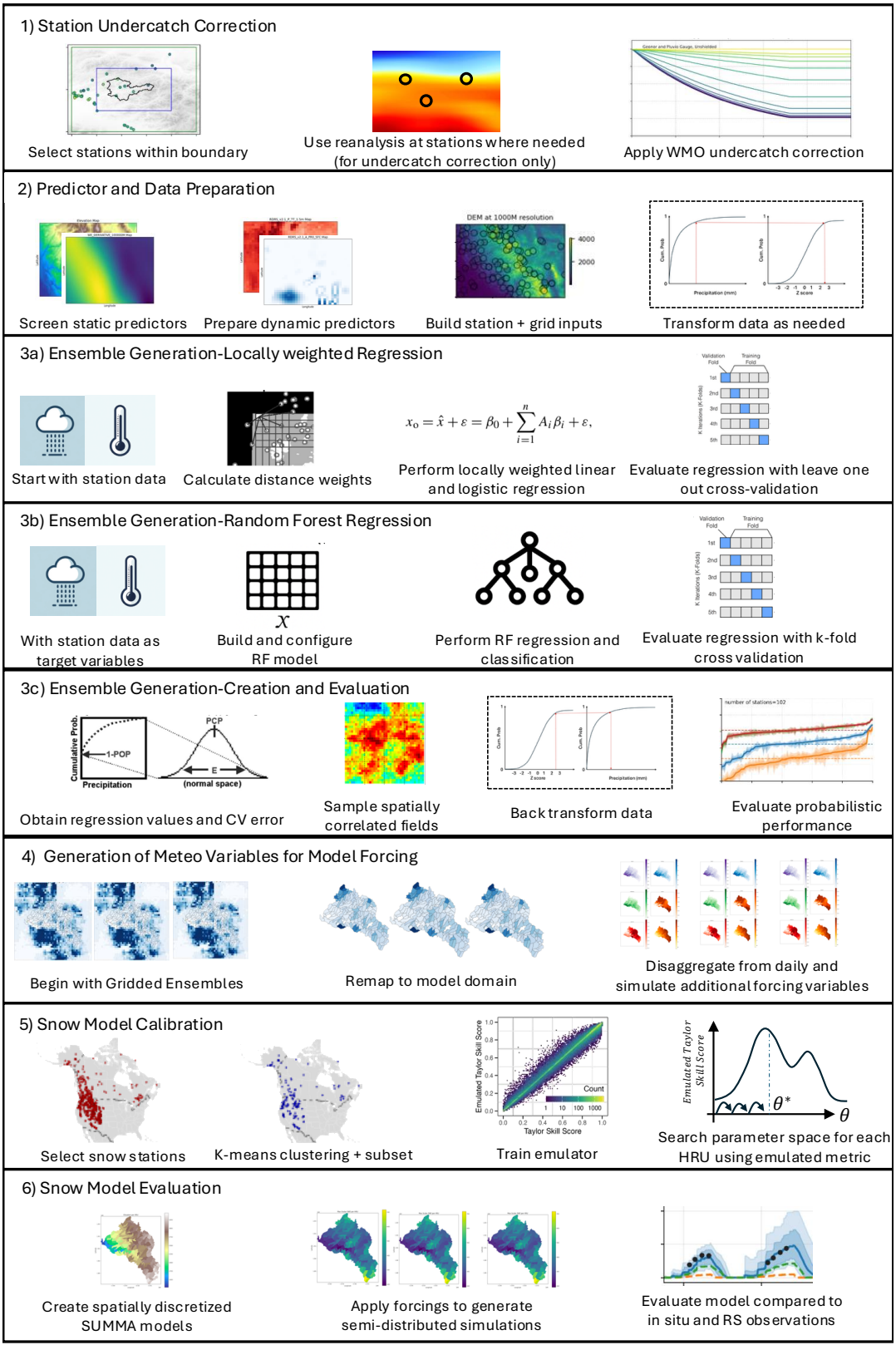


Figure 3. Overview of the Study Methodology. Methods 3-4 are components of the GPEP ensemble forcing generation approach.

The study methodology from data processing to model simulation is implemented with Snakemake workflows (Köster and Rahmann, 2012). Snakemake is a rule based management system that utilizes a Directed Acyclic Graph (DAG) to represent and manage the dependencies within a workflow. This allows reproducibility, transparency, and scalability across computational environments. An example DAG illustrating the workflow structure is provided in the supplementary material.

4.1 Station Undercatch Correction

Undercatch refers to the systematic underestimation of precipitation caused by wind effects, gauge configuration, and precipitation phase, which reduce the fraction of precipitation captured by instruments (Kochendorfer et al., 2017a, b). The bias is most pronounced for solid precipitation in cold and windy environments (Smith et al., 2022; Mekis and Vincent, 2011; Mekis et al., 2018), making undercatch correction an important step when generating precipitation reanalyses and datasets (Newman et al., 2020; Behrangi et al., 2018). A range of approaches exists for correcting undercatch, including empirical wind-based transfer functions, water-balance closure, and computational fluid-dynamics modeling (Colli et al., 2018, 2020; Beck et al., 2019). In this study, we apply empirical, gauge-based correction functions derived in the World Meteorological Organization (WMO) Solid Precipitation Intercomparison studies (Goodison et al., 1998; Kochendorfer et al., 2017b, a). These functions relate ~~collection~~ catch efficiency (CE) to wind speed and air temperature and are derived specific to gauge types and shielding configuration as shown in Figure 4.

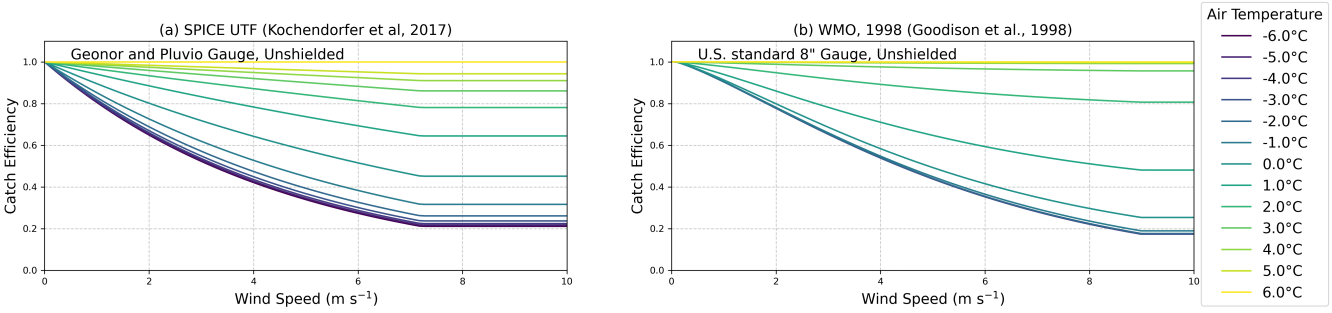


Figure 4. Catch efficiencies for the undercatch correction methods applied in this study based on temperature and wind speed. The updated SPICE UTF Kochendorfer et al. (2017a) to the Bow Basin is shown on the left side (a). On the right side (b) for the Chena and Tuolumne the approach from Goodison et al. (1998) and wind speed threshold behavior from Yang et al. (1998).

A regional approach that assumes consistent gauge type and configuration within each basin is taken. Specifically, in the Bow Basin the WMO SPICE UTF was applied for assumed unshielded Geonor and Pluvio gauges, and in the Chena and Tuolumne the WMO (1998) standard rain gauge correction is applied for assumed unshielded US standard 8" gauge (Goodison et al., 1998; Yang et al., 1998; Kochendorfer et al., 2017b, b). This follows a similar methodology to Newman et al. (2020) that derived large coverage station undercatch for Alaska, as well as undercatch corrections in Huidobro et al. (2023) that made similar assumptions due to lack of consistent gauge metadata across large domains of the United States. As ~~recognized in~~ demonstrated by Kochendorfer et al. (2017a) and Pierre et al. (2019), ~~who, while recognizing~~ along with the dominant effect

of shielding, ~~demonstrated~~ climate exerts a strong influence on catch efficiency relative to specific instrument type and the robustness of generalized transfer functions. The primary objective is not to ~~inter-~~compare methods directly, but to implement an appropriate correction for stations in each basin given available information. Due to a lack of station metadata in global datasets and with individual network providers, we assume a single gauge configuration per basin. Gridded reanalysis was used where ~~data were~~ co-located measurements were not available in an approach similar to Bertoncini and Pomeroy (2025). Wind observations at gauge height are used where available (e.g., GSOD), while ERA5 reanalysis winds translated to gauge height were used to supplement gaps for GHCN-D stations (Yang et al., 1998). ERA5 temperatures, lapsed to station elevation, are used to estimate precipitation phase and apply temperature-dependent correction factors. Full methodological details and equations are provided in Appendix A.

4.2 ~~Topographic Descriptor~~ Predictor and Data Preparation

In complex terrain, such as mountainous environments, topography is an important consideration to determine variables such as precipitation and temperature (Clark and Slater, 2006; Böhner and Antonić, 2009). Elevation impacts temperature through the adiabatic lapse rate, where temperature decreases with altitude (Daly et al., 2008). Precipitation patterns shift at higher elevations due to orographic lifting, leading to increased rainfall on windward slopes and rain shadows on leeward sides (Napoli et al., 2019). Additionally, slope and aspect determine the amount of solar radiation received, affecting localized heating and cooling rates (Böhner and Antonić, 2009). West-East and South-North elevation derivatives (gradients or slopes) capture directional changes in terrain, which can influence wind flow and the distribution of moisture, further impacting weather patterns (Robert et al., 2013).

Topographic descriptors are terrain variables calculated from a DEM to quantify elevation, spatial gradients and patterns. In this study, topographic descriptors derived from MERIT DEM are evaluated and applied as static predictor variables for precipitation and temperature. We compare simple intuitive descriptors of elevation, latitude, and longitude to derived ones, such as directional derivatives, Topographic Position Index (TPI), and aspect. Smoothing of topographic descriptors is performed to reduce noise and emphasize broader trends across the landscape as shown in Figure 5. This is explicitly examined by applying a 2D Gaussian filter with varying radii (200 m, 1,000 m, 10,000 m, 100,000 m, 150,000 m) to assess skill as a static predictor of precipitation and temperature. Most prior applications of GMET and GPEP have used elevation, latitude, longitude, and smoothed West-East and North-South elevation gradients as static predictors. ~~Optimal radii selected from a single predictor are subsequently used in a forward screening approach to build predictor sets for regression.~~

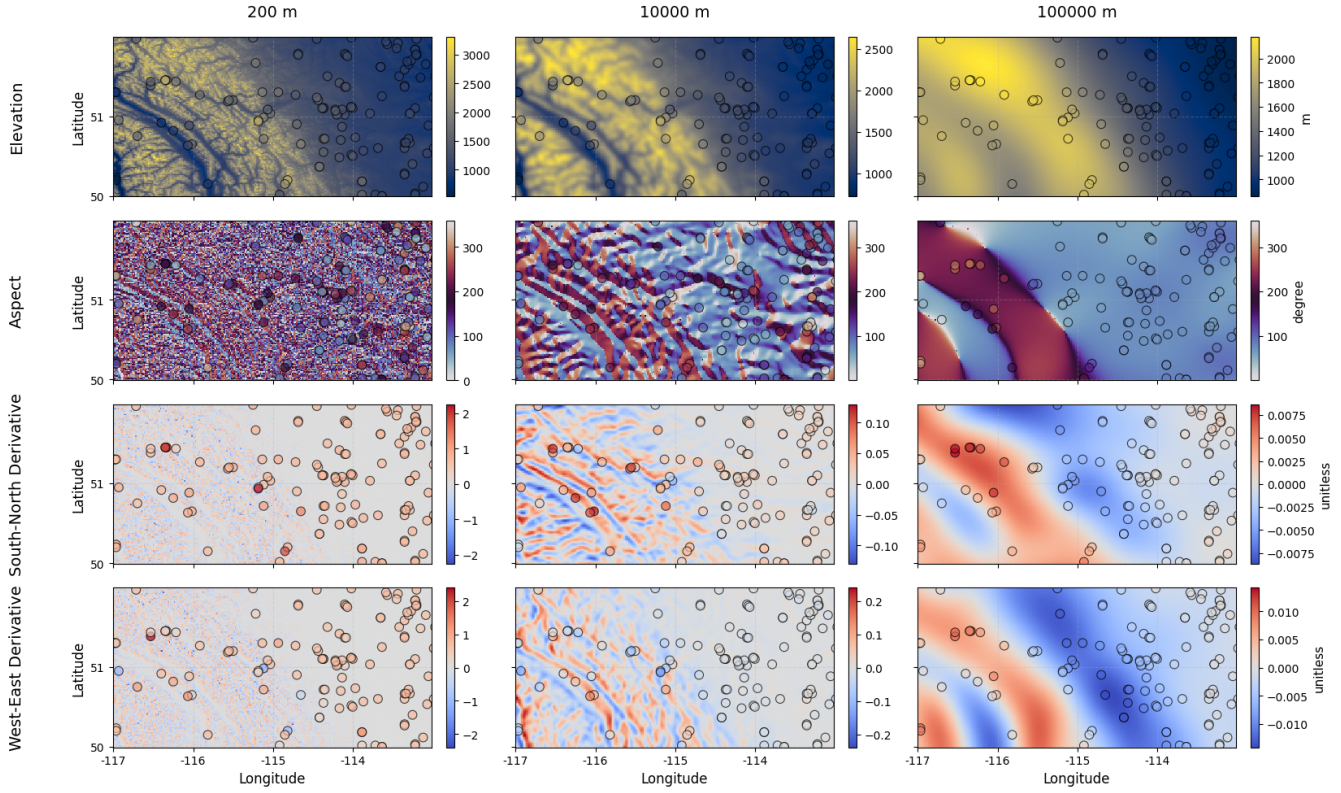


Figure 5. Visual comparison of topographic descriptors used as static predictors in deterministic regression. Calculated based on the MERIT-DEM, and applying a 2D Gaussian filter of varying radii (200 m, 10,000 m, and 100,000 m) examples for DEM, Aspect, South-North and West-East Derivatives are shown. Station location and values derived from the gridded products are indicated within the circles.

4.3 ~~Ensemble Meteorological Forcing Generation Approach~~

Predictor selection for the regression models used cross-validation error in a sequential forward screening procedure. An initial evaluation of static predictors was conducted to identify an appropriate level of spatial smoothing, after which the static variables were combined through iterative screening. Starting from an empty model, the static predictor producing the highest cross-validated Kling–Gupta Efficiency (KGE″) was first selected. Then additional static predictors were added one at a time, each time retaining the variable that yielded the greatest improvement in cross-validated KGE″. This process continued until the model skill no longer increased appreciably (defined as a gain in KGE″ < 0.01).

Dynamic predictors from gridded reanalysis were subsequently incorporated using the same forward screening procedure. The screening and regression fitting were performed independently for precipitation and temperature, allowing the most relevant static and dynamic predictors to be identified for each variable. The resulting predictor sets for precipitation and temperature were then combined to form the final predictor configuration used in the ensemble generation stage. The dynamic predictors consisted of the seven meteorological forcing variables required to drive the physically based hydrological model as

provided in the supplementary material. Precipitation and temperature were the primary variables of interest, while shortwave and longwave radiation, humidity, and wind speed were included to represent spatial variability in energy balance and atmospheric conditions. Together, these variables form the standard meteorological forcing set used in land surface models, ensuring that precipitation and temperature remain physically consistent with broader climatic drivers while maintaining practicality for operational implementation.

4.3 Ensemble Meteorological Forcing Generation

Meteorological datasets that represent spatially distributed meteorological fields are fundamental to hydrology and land-surface modeling. The fidelity of these datasets depends on their ability to capture spatial variability and to characterize associated uncertainties in observation-based estimates. To address this challenge, a probabilistic interpolation framework was first developed by Clark and Slater (2006), who implemented a Locally Weighted Regression (LWR) approach for generating ensembles of daily precipitation and temperature from station observations along with Spatially Correlated Random Fields (SCRF) to produce ensembles from sampling regression residuals. This foundational approach provides a statistical means of estimating both the mean and uncertainty of meteorological fields through spatial regression and stochastic simulation. Newman et al. (2015) refactored this approach into a formal ~~FORTRAN-based~~ Fortran-based software (the Gridded Meteorological Ensemble Tool, GMET) and produced a 100-member ensemble of daily precipitation and temperature for the contiguous United States (CONUS), demonstrating a practical capability to generate uncertainty-aware large-domain gridded forcings. Newman et al. (2020) applied the framework for Alaska, including station based undercatch based on regional gauge type assumptions. Subsequent enhancements enabled integration with deterministic gridded fields and data assimilation applications, providing dynamic uncertainty estimates through ensemble dressing (Liu et al., 2022). Finally, Bunn et al. (2022) advanced GMET by enabling the inclusion of dynamical predictors (e.g., from numerical weather prediction models, reanalyses or radar analyses), implementing formal cross-validation methods, and demonstrating improved performance in event-based precipitation estimation.

GMET was then generalized and re-implemented in the open-source Geospatial Probabilistic Estimation Package (GPEP; (Tang et al., 2024)), a Python-based platform extending ~~(and superseding)~~ GMET capabilities to enable treatment of a wider range of variables, predictor types, and spatial domains. GPEP now provides a modular framework for generating ensemble geophysical datasets (not only meteorological) using a variety of methods (the original regression and/or ML techniques), followed by stochastic sampling from spatially correlated random fields to produce ensemble realizations. The tool incorporates advances such as flexible predictor selection and station-based cross-validation for predictive uncertainty estimation. ~~For a more detailed description, the statistical formulation, parameter estimation, and ensemble generation processes are detailed in Appendix B, as well as the configuration choices specific to this implementation. Ensembles were generated at a $1/64^{\circ}$ resolution, or approximately 1 km grid spacing.~~

In this study, we evaluate the ~~Random Forest~~ random forest (RF) capability within GPEP and compare its performance with the original locally-weighted regression (LWR) approach. Both methods are systematically applied to assess how individual and combined topographic predictors improve geospatial estimation in mountainous terrain. Ensembles were generated at a

1/64° resolution, or approximately 1 km grid spacing at a daily timestep. We further examine the contribution of dynamic meteorological grids from reanalyses to better capture transient meteorological variability. By linking station observations with gridded topographic and reanalysis data, this approach enables an explicit evaluation of how terrain features and large-scale meteorological conditions jointly influence skill. Together, these experiments provide insight into the relative strengths of machine-learning and regression-based approaches for probabilistic meteorological estimation in complex terrain.

4.3.1 ~~Regression Cross Validation~~

A simple schematic overview of the approach is provided in Figure 3. For a more detailed description, the statistical formulation, parameter estimation, and ensemble generation processes are detailed in Appendix B, as well as the configuration choices specific to this implementation.

4.3.1 Regression Cross-Validation

Cross validation results measure the regression error at station locations. In this study, leave-one-out cross-validation is used for locally weighted regression, and k-fold cross-validation (k=10) is used for ~~Random Forest~~random forest. Involving the division of the dataset into k subsets (k=10), with the model trained on k-1 subsets and validated on the remaining one. This process is repeated k times, each time with a different subset serving as the validation set.

GPEP offers 16 metrics for the regression cross-validation that can be used to evaluate regression approaches. As in Tang et al. (2024), an augmented Kling-Gupta Efficiency, KGE″, is used because it normalizes the bias term using the standard deviation instead of the mean, making it more appropriate for temperature variables. This formulation mitigates the influence of measurement units (e.g., Kelvin versus degrees Celsius) and reduces the exaggerated bias that can occur near zero degrees Celsius (Tang et al., 2021). Precipitation values are evaluated in transformed form (e.g. with Box-Cox). See also Santos et al. (2018) for a discussion regarding near zero values on transformed values of streamflow.

$$\mathrm{KGE}'' = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + \beta^2}, \tag{1}$$

where r is the Pearson correlation coefficient between observations (o) and estimates (e), $\alpha = \frac{\sigma_e}{\sigma_o}$ is the variability ratio, $\beta = \frac{\mu_e - \mu_o}{\sigma_o}$ is the bias term, μ is the mean value, and σ is the standard deviation.

4.3.2 ~~Predictor Screening~~ Probabilistic Skill Evaluation and ~~Combination Building~~Bootstrapping

~~Predictor selection for the regression models used cross-validation error in a sequential forward screening procedure. An initial evaluation of static predictors was conducted to identify an appropriate level of spatial smoothing, after which the static variables were combined through iterative screening. Starting from an empty model, the static predictor producing the highest cross-validated Kling–Gupta Efficiency (KGE″) was first selected. Then additional static predictors were added one at a time, each time retaining the variable that yielded the greatest improvement in cross-validated KGE″. This process continued until the model skill no longer increased appreciably (defined as a gain in KGE″ $\leq$ 0.01).~~

Dynamic predictors from gridded reanalysis were subsequently incorporated using the same forward screening procedure. The screening and regression fitting were performed independently for precipitation and temperature, allowing the most relevant static and dynamic predictors to be identified for each variable. The resulting predictor sets for precipitation and temperature were then combined to form the final predictor configuration used in the ensemble generation stage. The dynamic predictors consisted of the seven meteorological forcing variables required to drive the physically based hydrological model. Precipitation and temperature were the primary variables of interest, while shortwave and longwave radiation, humidity, and wind speed were included to represent spatial variability in energy balance and atmospheric conditions. Together, these variables form the standard meteorological forcing set used in land surface models, ensuring that precipitation and temperature remain physically consistent with broader climatic drivers while maintaining practicality for operational implementation. A complete list and description of the dynamic forcing variables used in this study are provided in Appendix **??**.

4.4 Probabilistic Skill Evaluation and Bootstrapping

To evaluate the probabilistic performance of the ensemble meteorological estimates against station observations, we use the Brier Skill Score (BSS). The BSS measures the relative skill of a probabilistic prediction compared to a reference prediction and is widely applied in ensemble meteorology and hydrology to quantify the reliability and resolution of probabilistic predictions (Wilks, 2011; Hallouin et al., 2024).

The BSS is defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{ens}}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{2}$$

where $\mathrm{BS}_{\mathrm{ens}}$ is the Brier Score (BS) of the ensemble prediction and $\mathrm{BS}_{\mathrm{ref}}$ is the Brier Score of a reference prediction, in this study the observed climatology. A BSS of 1 indicates perfect probabilistic prediction, a value of 0 indicates no improvement over the reference, and negative values indicate degraded skill.

The underlying Brier Score (BS) represents the mean squared error between the predicted probability f_i of an event (e.g., precipitation occurrence or exceedance of a threshold) and the corresponding binary observation o_i (1 if the event occurred, 0 otherwise):

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{3}$$

where N is the number of predictions.

In this study, the ensemble probability of event occurrence (e.g., daily precipitation $> 0\ \mathrm{mm}$ is computed from the fraction of ensemble members predicting the event at each station location (Clark et al., 2006). The resulting probabilities are compared against observed station data to calculate the BSS, providing a quantitative measure of how well the ensemble captures the observed event frequency and variability at station locations.

To quantify the sampling uncertainty of the Brier Skill Score, we apply a nonparametric resampling bootstrap method, following the approach of Clark et al. (2021a) as implemented in Hallouin et al. (2024). Bootstrapping provides an intuitive and data-driven means to estimate the empirical probability distribution of performance statistics, enabling confidence or tolerance intervals to be derived without reliance on analytical assumptions. Specifically, we employ the non-overlapping block bootstrap (NBB) method which resamples data in one-year blocks (Carlstein, 1986). For each bootstrap replicate, the BSS is recomputed, producing an ensemble of estimates from which the tolerance intervals are determined.

4.4 ~~Probabilistic Physically Based Snow Modeling~~Generation of Meteorological Variables for Model Forcing

Physically based hydrological models require a wider range of meteorological inputs and finer temporal resolution than conceptual models because they explicitly represent both the energy and water balances. Sub-daily forcing is needed to resolve diurnal variations in key variables such as temperature and incoming solar radiation, which control melt and phase-change processes. These additional forcings enable a more detailed and physically consistent simulation of hydrological processes compared to conceptual approaches, where empirical relationships such as degree-day methods govern fluxes like snowmelt (Mosier et al., 2016). However, many of the required variables, including shortwave radiation, are not routinely observed at the spatial and temporal resolution needed to drive distributed hydrological models. In this study, such variables are therefore simulated using precipitation, temperature, location and topographic information. As depicted in Figure 3, gridded data is first remapped to the model sub-basins, disaggregation from daily data and simulation of additional variables follows.

MetSim (short for Meteorology Simulator), is a Python package for the estimation of meteorological variables to support hydrological and land surface modeling (Bennett et al., 2020). The core algorithm for MetSim MTCLIM (v3.4) contains a mountain microclimate simulation model, three main computation modules for solar geometry, meteorological simulation, and temporal disaggregation (Hungerford et al., 1989). The MTCLIM meteorological downscaling scheme derives additional climate variables from daily minimum and maximum temperature and precipitation for use in hydrologic and land–surface models. Daylight mean temperature is estimated as a weighted average of extremes (Thornton et al., 1997), while cloud transmittance and cover are inferred from the relationship between diurnal temperature range, its smoothed climatology, and precipitation occurrence (Kimball et al., 1997). Dew point temperature and vapor pressure are obtained iteratively in conjunction with shortwave radiation and potential evapotranspiration, ensuring consistency among humidity, radiation, and evaporative demand. Shortwave radiation is calculated from potential clear-sky radiation, transmissivity, and humidity corrections (Bohn et al., 2013). More detailed algorithm description and implementation are provided in Bennett et al. (2020) and Bohn et al. (2013) .

4.5 Snow Modeling

4.5.1 Model Description

The SUMMA hydrological modeling framework is used in this study for snow modeling. It provides physical representations of snow accumulation, interception, metamorphism, and ablation processes using an energy balance approach to calculate

temperature and phase change (Clark et al., 2015b, c, 2021b). The snowpack is represented as a dynamic set of layers, each with specific physical properties, allowing detailed simulation of the snowpack, including compaction, sublimation, meltwater, and refreezing (Clark et al., 2015a).

SUMMA was selected because it offers a comprehensive, process-based framework that extends beyond snowpack modeling to hydrological simulation, enabling future translation of snow states and ensemble uncertainties into runoff predictions. Its modular design and explicit energy-balance formulation make it particularly suited to evaluate the effects of meteorological forcing uncertainty on snowpack evolution and hydrological response. While SUMMA does not explicitly represent wind redistribution or avalanching, which remain limitations for steep alpine terrain, it provides a physical basis and extensible framework for quantifying how forcing uncertainty propagates through snow accumulation and melt processes to influence basin runoff. SUMMA model decisions in this study, which control parameterization and physics descriptions in the model framework are consistent with Farahani et al. (2025). With respect to snow modeling, the model decisions select the dynamic snowpack layering scheme adapted from the Community Land Model Lawrence et al. (2011). Model decisions also inform snowpack boundary conditions, such as the throughfall of snow based on new snow density, itself formulated as an empirical function of temperature Hedstrom and Pomeroy (1998). Snow density is also used to calculate thermal conductivity Jordan (1991). The albedo decay rate is held constant in time Verseghy (1991). For greater detail on the snow model implementation and model decisions, the reader is referred to (Clark et al., 2015a).

The Chena, Bow, and Tuolumne River Basins in this study are discretized to represent the watershed as a collection of Grouped Response Units (GRUs). While this study focused on snow, in fact the hydrological simulation extends to runoff generation. In this approach, a vertical column for each GRU is conceptualized to model the movement and transformation of water and energy through different layers, from the vegetation canopy down to the soil and underlying groundwater. This column-based representation allows for detailed simulation of processes such as interception, infiltration, soil moisture dynamics, evapotranspiration, and groundwater recharge. The semi-distributed nature of the model helps to represent essential spatial variability and dominant controls over land surface hydrology. TDX-Hydro is used directly for the spatial discretization of GRUs in this study. ~~While Hydrologic Response Units (HRUs) can be derived from GRUs to capture finer spatial variability, this study focuses solely on GRUs.~~ This is due to the high level of discretization in the TDX-Hydro dataset, and for clarity in evaluation. This approach captures elevation most directly in mountainous watersheds, without explicitly including aspect or slope. Precipitation and temperature ensemble gridded data are remapped to the GRUs defined by TDX-Hydro vector data using the EArth SYstem MOdeling REmapper (EASYMORE) package (Gharari et al., 2023). This remapping is a spatial average of gridded data calculated over the GRU units.

Each GRU is parameterized using globally consistent geospatial datasets following the standardized workflow of Knoben et al. (2022). Elevation is derived from the MERIT Hydro hydrologically adjusted DEM (Yamazaki et al., 2019), ensuring accurate flow connectivity and representation of topographic gradients. Soil properties are obtained from SoilGrids 250 m (Hengl et al., 2017), where sand, silt, and clay fractions are converted to USDA texture classes and the dominant soil type is assigned to each GRU. Land cover is taken from MODIS MCD12Q1 v6 (Friedl and Sulla-Menashe, 2019) using the IGBP classification to define vegetation type and surface characteristics. These layers are spatially intersected with the GRU geometries to

extract mean elevation and modal soil and land-cover classes, forming the static physiographic attributes used by SUMMA to parameterize vegetation, surface roughness, and soil hydraulic and thermal properties.

4.5.2 ~~Simulation of Meteorological Variables for Model Forcing~~

~~Physically based hydrological models require a wider range of meteorological inputs and finer temporal resolution than conceptual models because they explicitly represent both the energy and water balances. Sub-daily forcing is needed to resolve diurnal variations in key variables such as temperature and incoming solar radiation, which control melt and phase-change processes. These additional forcings enable a more detailed and physically consistent simulation of hydrological processes compared to conceptual approaches, where empirical relationships such as degree-day methods govern fluxes like snowmelt (Mosier et al., 2016). However, many of the required variables, including shortwave radiation, are not routinely observed at the spatial and temporal resolution needed to drive distributed hydrological models. In this study, such variables are therefore simulated using precipitation, temperature, location and topographic information.~~

~~MetSim (short for Meterology Simulator), is a Python package for the estimation of meteorological variables to support hydrological and land surface modeling (Bennett et al., 2020). The core algorithm for MetSim MTCLIM (v3.4) contains a mountain microclimate simulation model, three main computation modules for solar geometry, meteorological simulation, and temporal disaggregation (Hungerford et al., 1989). The MTCLIM meteorological downscaling scheme derives additional climate variables from daily minimum and maximum temperature and precipitation for use in hydrologic and land–surface models. Daylight mean temperature is estimated as a weighted average of extremes (Thornton et al., 1997), while cloud transmittance and cover are inferred from the relationship between diurnal temperature range, its smoothed climatology, and precipitation occurrence (Kimball et al., 1997). Dew point temperature and vapor pressure are obtained iteratively in conjunction with shortwave radiation and potential evapotranspiration, ensuring consistency among humidity, radiation, and evaporative demand. Shortwave radiation is calculated from potential clear-sky radiation, transmissivity, and humidity corrections (Bohn et al., 2013). More detailed algorithm description and implementation are provided in Bennett et al. (2020) and Bohn et al. (2013) .~~

4.5.2 Snow Model Calibration

We calibrate eleven SUMMA parameters related to snow processes using the Large Sample Emulator (LSE) approach introduced in Tang et al. (2025b) for the Community Land Model and Farahani et al. (2025) for SUMMA. The LSE jointly leverages model simulations from a large sample of locations by using ML techniques to emulate a comprehensive model response surface, and uses the emulator as an inexpensive parameter set search tool in iterative sequential optimization.

First, for each location (e.g., snow gauges), we run SUMMA using $N_{i=0}$ parameter sets. These parameter sets are obtained by using Latin Hypercube Sampling, and the parameter values vary across sites. For each parameter set, we compute performance metrics (e.g., Root Mean Square Error, Pearson correlation coefficient) by contrasting observed and simulated SWE values. ~~Second,~~

Second, to estimate the surface response, we train an emulator of the simulated surface response using parameter sets and climate attributes from each location as predictors (x), and the performance metric of interest as the predicted value (y). To this end, we use a matrix of $N_{sites} \cdot N_{i=0}$ rows and $N_\theta + N_{Atts}$ columns as x, where N_θ is the number of parameters to calibrate, and N_{Atts} the number of site attributes. Note that, for each site, the attributes remain constant among rows. y has $N_{sites} \cdot N_{i=0}$ rows and only one column. Then, the emulator relates $y = g(x) + \epsilon$, where g is the emulator (e.g., Random Forest) and ϵ represents the error.

Third, we utilize the trained emulator to search for emulated surface responses and identify regions of the parameter space that may lead to improved performance metrics. The search is conducted at each site (the same one used for training) by using the Genetic Algorithm (GA) with a total budget of 10,000 runs (note that, in this step, the emulator is used in the GA search, not SUMMA). For each site, we select the parameter sets associated with the best 100 emulated responses and run the SUMMA model to obtain the actual performance metric. This leads to $N_i \cdot N_{site}$ new performance metric values, with $N_i = 100$. Finally, the new values are concatenated into the original matrix (now with $N_{sites} \cdot (N_{i=0} + N_i)$ rows), and a new emulator can be trained. This process is repeated until some criteria are met. In our case, we iterate until each site has 1,000 SUMMA runs (i.e., five iterations, starting with $N_{i=0} = 500$ parameter sets). We use the Taylor Skill Score (TSS) from Taylor (2001) as a target metric since it comprises temporal correlation and variance.

~~We~~ To train the emulator, we consider snow gauges across Canada and the US, with four water years of measurements, each with at least 20 measurements per year taken from NorSWE dataset and quality-controlled SNOTEL stations (770 stations in total~~) (Sun et al., 2019; Mortimer and Vionnet, 2025)~~; the location of the stations and the emulator error are shown in Figure C1 in Appendix C; Sun et al. 2019; Mortimer and Vionnet 2025). To reduce the computational burden, we conduct K-means clustering using climate attributes. For each resulting cluster, we randomly select 20 stations. If a cluster has fewer than 20 stations, all stations are considered to preserve equal representation from all clusters. Thus, to train the emulator, we reduce the number of snow gauges to 162. Note that for each of these stations, we obtain the required forcing data to run SUMMA from the ERA5 reanalysis, as well as the SWE measurements to compute performance metrics. Hence, the final LSE comprises 162,000 SUMMA runs (162 stations · 1,000 parameter sets per station). Note that for each of these stations, we obtain the required forcing data to run SUMMA from the ERA5. This LSE is subsequently used to obtain parameter sets at each model GRU by searching the emulated surface response (which is conditional on the GRU's local geo-attributes) using the GA. Note that in this case, only one search is conducted (i.e., no iteration is carried out) and the parameter set associated with the best emulated performance metric is assigned to each GRU. Parameter ranges used in this calibration are provided in ~~Appendix ??~~the supplementary material.

4.6 Snow Model Evaluation~~Using In-situ Measurements~~

4.6.1 Evaluation with in situ measurements

Evaluation of the snow simulations focuses on comparing modeled SWE from SUMMA against ~~in-situ~~ in situ snow observations. SUMMA is run with each meteorological ensemble member, with input gridded data spatially averaged over each model

GRU (Figure 2). Model simulations are compared to manual snow survey measurements from the NorSWE dataset, which provide estimates of SWE across representative terrain transects. These surveys are selected for model evaluation because they better capture local variability than automated snow pillows, which record conditions at a single point. Snow pillows also record conditions in the absence of overhead forest canopy, which may influence the representativeness of comparisons with spatially averaged model outputs.

Model performance is assessed using skill target diagrams (Mudryk et al., 2025; Jolliff et al., 2009), which quantify model skill in terms of both temporal pattern similarity and bias. The total skill metric is expressed as the normalized two–component distance. The method uses a normalized spatial pattern (S_{pattern}) and bias (S_{bias}) to quantify model performance (S_{total}):

$$S_{\text{total}} = \sqrt{S_{\text{pattern}}^2 + S_{\text{bias}}^2} \tag{4}$$

$$S_{\text{pattern}} = f \cdot \left[1 - \frac{2(1+R)}{(\sigma_* + 1/\sigma_*)^2}\right] \tag{5}$$

$$S_{\text{bias}} = f \cdot \frac{|b|}{\text{uRMSE}_{\text{max}}} \|S_{\text{pattern}}\|_{\text{max}} \tag{6}$$

In these expressions, b is the mean bias, $\text{uRMSE}_{\text{max}}$ is the maximum unbiased RMSE among products (in absolute units), $\|S_{\text{pattern}}\|_{\text{max}}$ scales bias to the range of pattern errors, and f is a cross-test scaling factor set to 1 in this study. This standard relationship among uRMSE, R, and σ_* is also represented using Taylor diagrams to summarize multiple aspects of model performance for the snow simulations (Taylor, 2001). The standard-deviation ratio and its relationship to normalized unbiased RMSE are given by:

$$\sigma_* = \frac{\sigma_x}{\sigma_r} \tag{7}$$

$$\text{uRMSE}_*^2 = 1 + \sigma_*^2 - 2\sigma_* R \tag{8}$$

4.6.2 Evaluation of Remotely Sensed Data

MODIS-derived Snow Cover Fraction (SCF) and Snow Disappearance Date (SDD) were computed following the general workflow and approach of Crumley et al. (2020). The workflow was adapted to compute zonal statistics over hydrological sub-basins rather than point locations. SCF was calculated as the ratio of snow-covered to valid MODIS observations within each water year, while SDD was estimated using a persistence-based transition criterion defined as the latest snow-covered day followed by at least five valid snow-free observations within a 15-day period. Annual SCF and SDD metrics were summarized for each sub-basin and aggregated basin-scale comparison per water year.

Airborne Snow Observatory (ASO) SWE rasters were aggregated to the model sub-basins to allow comparison with modelled sub-basin SWE and in situ data. This provided an independent assessment of seasonal evolution of simulated snow accumulation for the Tuolumne from 2014-2018.

4.6.3 Evaluation of Precipitation Phase Uncertainty

In SUMMA, precipitation phase is determined using wet bulb temperature-based partitioning governed by calibrated snow parameters, most notably tempCritRain and tempRangeTimestep. These parameters control the central rain–snow threshold and the width of the transition range over which precipitation is partitioned between liquid and solid forms. As a result, uncertainty in air temperature near freezing directly propagates into uncertainty in precipitation phase and snowfall quantity

Meteorological forcing uncertainty was also evaluated by quantifying its influence on precipitation phase partitioning and seasonal snowfall accumulation. Snowfall fraction was calculated as the fraction of total precipitation occurring as snowfall, excluding time steps with no precipitation. Snowfall fraction was averaged across all GRUs within each catchment and composited by day of the water year (1 October–30 September) to produce a climatological annual cycle.

To characterize the spatial distribution of precipitation phase uncertainty, total water-year snowfall accumulation was calculated for every GRU and ensemble member. The ensemble standard deviation of water-year snowfall accumulation was then computed for each GRU as a measure of the absolute uncertainty in seasonal snowfall. To facilitate comparison across basins with differing snowfall totals, the standard deviation was also normalized by the ensemble mean water-year snowfall accumulation, yielding a dimensionless measure of relative uncertainty.

5 Results

5.1 Undercatch Correction

Catch Efficiency (CE) is the ratio of recorded to actual precipitation with lower CE values corresponding to greater undercatch correction. Figure 6 compares aggregated CE values calculated in this study for each basin throughout the year, as well as the net effect on accumulated precipitation over the water year. The seasonal and spatial patterns remain consistent with those reported in detailed field studies (Smith et al., 2022).

The use of reanalysis wind estimates to enable empirical correction of station undercatch was a key step, as indicated in Figure 6 wind gap filling stations. Reanalysis wind underestimated high wind speeds compared to observations resulting in conservative CE estimates. ~~This~~ These biases can be improved by using daily maximum values from reanalysis, matching physical process understanding that snowfall events occur during storm periods. The use of ERA5 daily maximum wind speeds yields conservative CE estimates with other variables held constant reflects the tendency of reanalysis winds to underestimate site-level exposure (Kochendorfer et al., 2017b).

Further detail on the methodological choices and comparisons with reference datasets is provided in Appendix A. Although the approach requires several simplifying assumptions, it effectively captures and corrects wind-induced undercatch, an established and influential source of bias in mountainous snowfall and precipitation measurements.

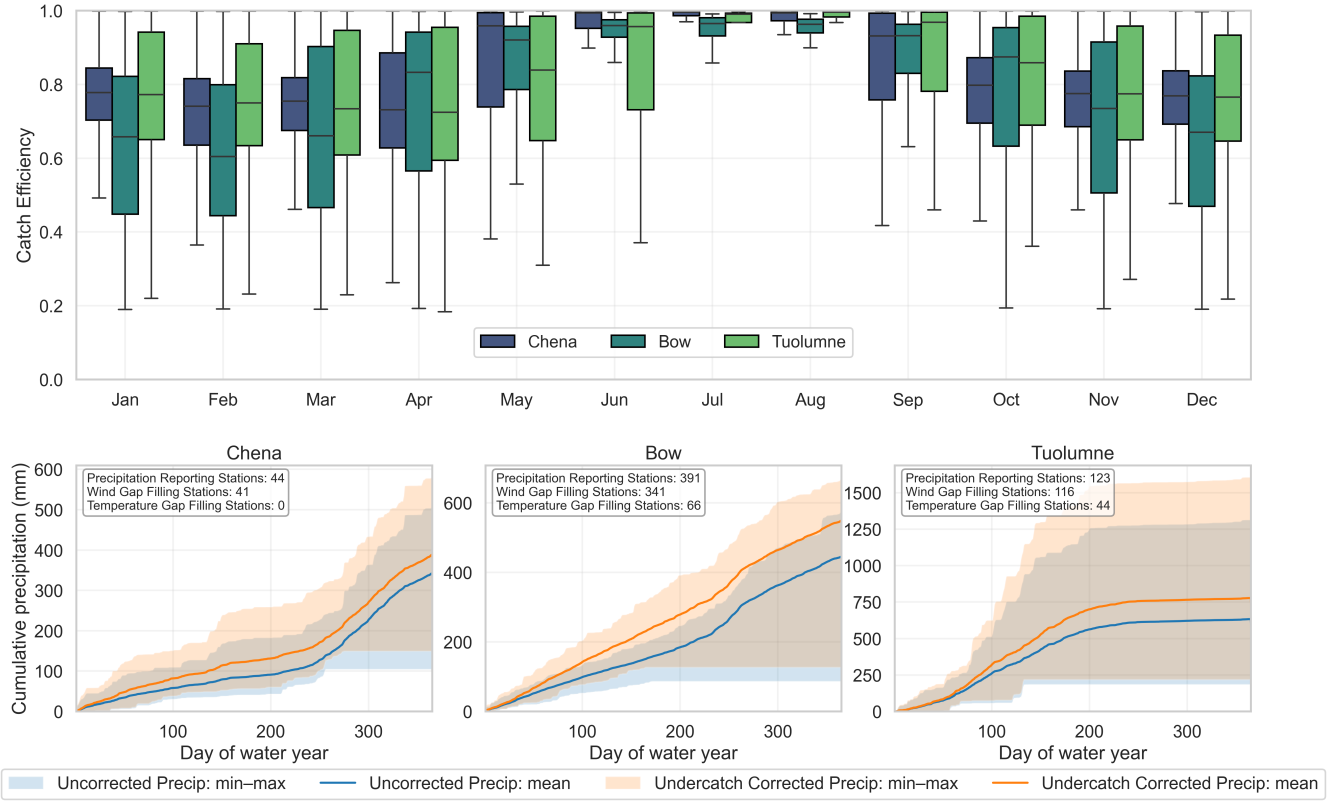


Figure 6. Gauge catch efficiency (CE) and ~~Cumulative Precipitation~~cumulative precipitation across basins. The top panel shows CE correction ranges ~~per~~ by month across the Bow, Chena, and Tuolumne basins. The bottom panel shows effect of undercatch correction on precipitation accumulation over the water year (starting Oct 1st). Each water year and the mean of all years and stations is shown for each basin. Both the boxplots and area time series summarize calculations at station locations grouped by basin

5.2 Static Predictor Screening with Cross Validation

The initial screening of static topographic predictors used leave-one-out cross-validation (LWR) and 10-fold cross-validation (RF) at station locations. For each basin, single predictors were evaluated individually, with performance summarized using cross-validated KGE″ for daily precipitation (in transformed space) and daily mean temperature. Figure 7 shows the results for the Bow, with analogous figures for the Chena and Tuolumne provided in the Supplement. For single predictors, LWR typically yields slightly higher KGE″ than RF, reflecting the advantage of local weighting when only one static covariate is used.

For temperature, elevation and smoothed DEM fields are consistently skillful predictors for both LWR and RF, reflecting the dominant lapse-rate control on air temperature. Applying spatial smoothing with a larger Gaussian radius (more smoothing) to derived topographic descriptors (e.g., slope, aspect, terrain derivatives) improves cross-validation skill, as smoothing suppresses small-scale noise while retaining large-scale orographic gradients.

For precipitation, the relative performance of individual static predictors is more even. Smoothed variants of terrain de-
605 scriptors (e.g., South-North and West-East) generally outperform their unsmoothed forms. Across all three basins, predictors smoothed with greater Gaussian radii perform better, indicating that large-scale physiographic structure is more informative than fine-scale terrain details for daily precipitation estimation at the available station density.

Based on the screening results, topographic descriptors smoothed with 100 km Gaussian radius were retained for subsequent predictor combination building, along with unsmoothed elevation, latitude, and longitude.

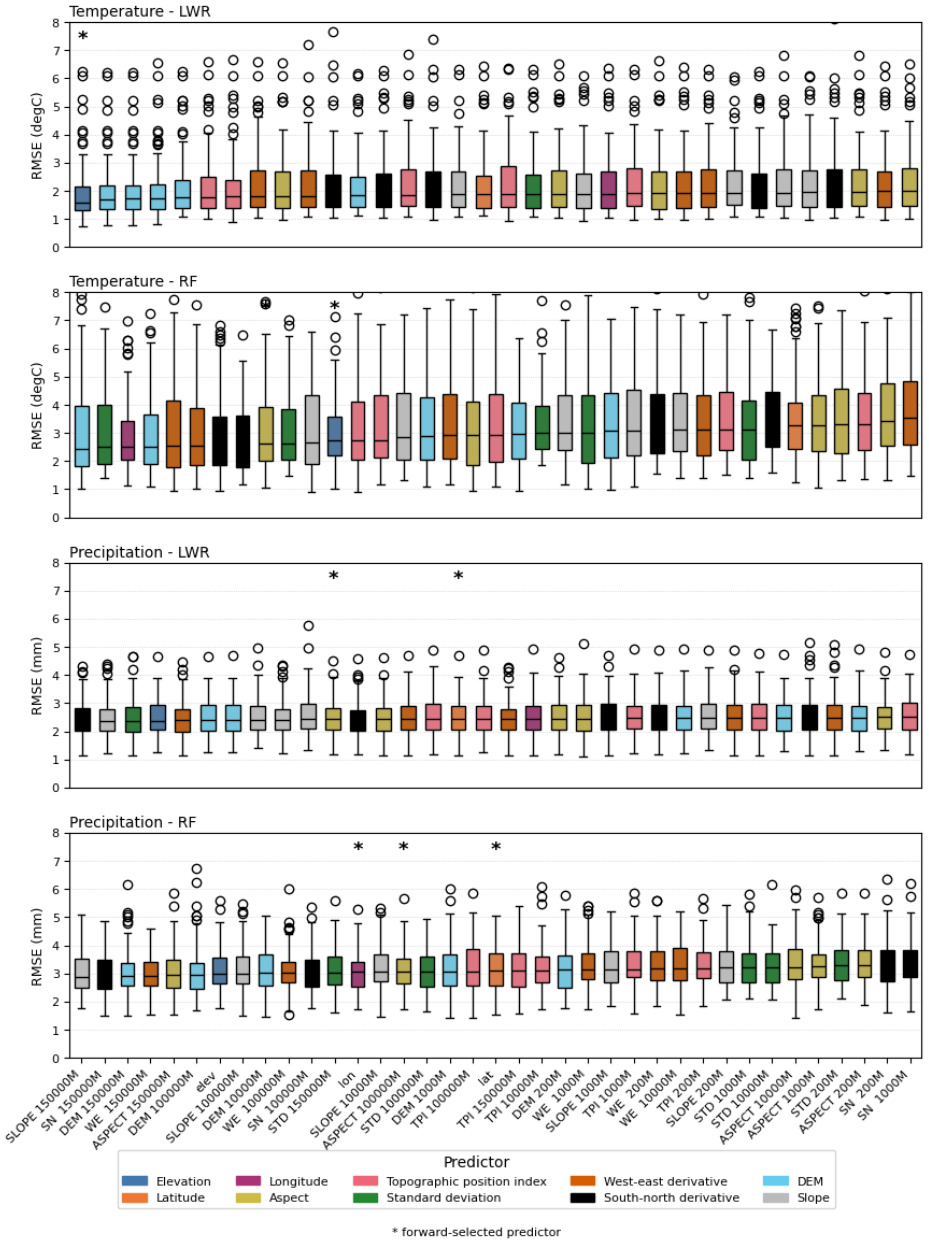


Figure 7. Static predictor screening showing single predictor cross validation results for the Bow. Topographic descriptors and their smoothed variants (200 m, 1,000 m, 10,000 m, 100,000 m, 150,000 m) are evaluated to assess cross-validation RMSE for Temperature and Precipitation for both Locally Weighted Regression (LWR) and Random Forecast (RF). Asterisk indicates the predictor retained in sequential forward screening.

5.3 Predictor Combination Selection and Regression Performance

Sequential forward screening was used to construct multivariate predictor sets separately for precipitation and temperature, and for each basin and regression method. Starting from an empty model, the static predictor that maximized cross-validated KGE″ was selected, after which additional predictors were added one at a time. Predictors were retained only if they increased KGE″ by more than 0.01, and the procedure was repeated after allowing dynamic reanalysis-based variables to enter the pool. The resulting optimal combinations are summarized in Table 1.

Table 1. Best static and static+dynamic predictor combinations by catchment, variable, and model (LWR - Locally Weighted Regression, RF - Random Forest).

Catchment	Variable	Model	Static predictors	+ Dynamic predictors
Chena	Precipitation	LWR	South–North Derivative	Surface pressure
		RF	Elevation, Longitude	Precipitation rate
	Temperature	LWR	Elevation	Air temperature
		RF	Latitude, Elevation, South–North Derivative	Humidity
Bow	Precipitation	LWR	Aspect, Latitude	Air temperature
		RF	Aspect, Latitude, Longitude	Surface pressure
	Temperature	LWR	Elevation	Shortwave radiation
		RF	Aspect, Elevation	Air temperature
Tuolumne	Precipitation	LWR	Elevation, West–East Derivative, Latitude	Air temperature
		RF	Aspect, Elevation, Latitude, Longitude	Precipitation rate
	Temperature	LWR	Elevation	Air temperature
		RF	Elevation, South–North Derivative	Air temperature

Across the study basins, elevation or its smoothed derivatives ~~dominate~~ are the primary controls on temperature estimation, with only a small number of additional predictors ~~needed~~contributing. For LWR, elevation alone ~~provides~~ explains most of the ~~skill, and~~ variability, while adding a single dynamic predictor (such as reanalysis air temperature or shortwave radiation~~from reanalysis) yields marginal but systematic~~) provides modest but consistent improvements. RF models for temperature ~~typically select~~ use combinations of elevation and one or two additional topographic descriptors (e.g., aspect ~~, south–north~~ or south–north derivative), ~~then further improve performance~~ with further gains when reanalysis air temperature or humidity is included. For precipitation, both static and dynamic predictors ~~are important~~contribute. In the Bow ~~, LWR favors~~ basin, LWR identifies aspect and latitude as key predictors, with air temperature as the ~~dominant dynamic addition, whereas RF also selects~~ main dynamic input, while RF additionally includes longitude and surface pressure. In the Chena and Tuolumne ~~, RF tends to use~~ basins, RF uses elevation and horizontal derivatives ~~in combination~~ together with reanalysis precipitation rate and surface pressure,

~~suggesting that non-linear~~ indicating that interactions between orography and large-scale meteorology ~~are important for daily precipitation~~ play an important role in daily precipitation patterns within these basins.

Figure 8 shows the distribution of cross-validated KGE″ across all stations for the main predictor combinations. RF systematically outperforms LWR once multiple predictors are available, particularly for precipitation. Adding dynamic predictors further shifts the cumulative distributions to the right, indicating higher skill for the majority of stations. Temperature KGE″ is relatively high and similar across methods, whereas precipitation shows a clearer separation between LWR and RF, and between static-only and static+dynamic configurations. It is notable that inclusion of dynamic predictors can degrade the skill of LWR predictions, potentially due to overfitting or creating localized patterns that are not representative.

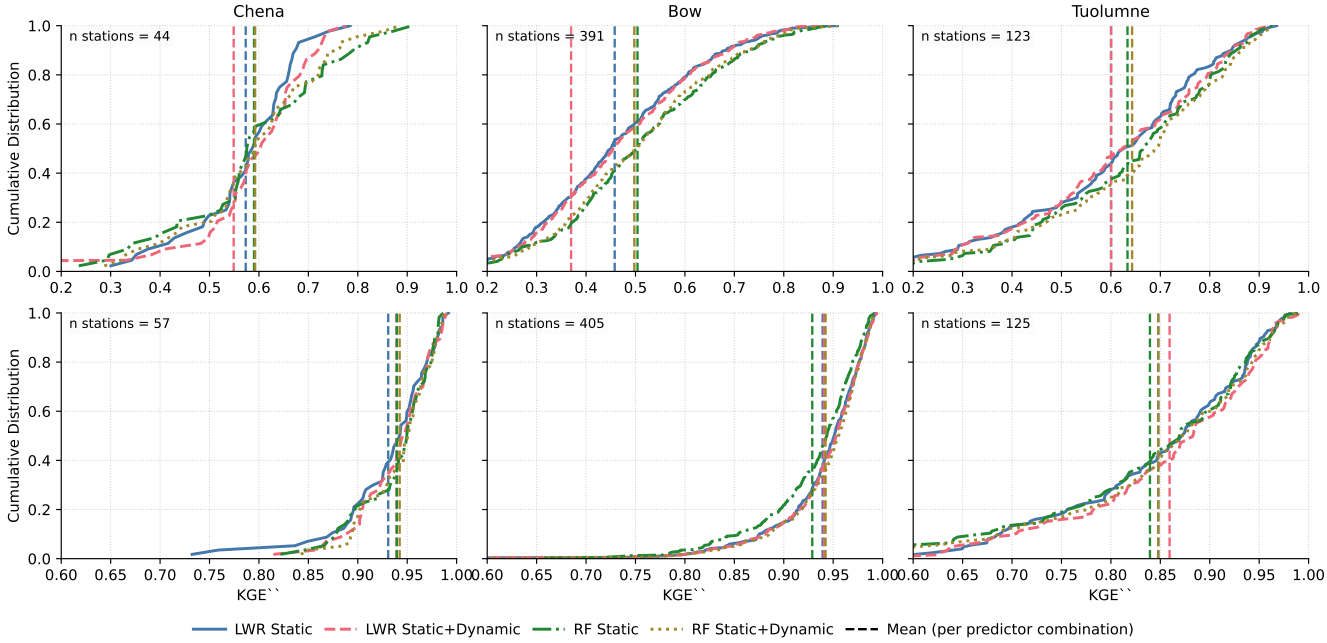


Figure 8. Cross validation KGE″ values for predictor combinations from forward screens. The KGE″ for all stations are shown as a Cumulative Distribution Function. Results further to the right of the x-axis are better. The number of stations available for each analysis is shown.

The spatial pattern of cross-validation performance for the RF static+dynamic combinations is shown in Figure 9. There is no strong dependence of KGE″ on elevation, station density, or location within each basin. ~~Most stations~~ Stations achieve KGE″ > 0 for precipitation and KGE″ > 0.4 for temperature, with ~~a small subset of outliers~~ the proportion of stations exhibiting lower skill not exceeding 3% for any catchment+variable combination and being 0% in most cases, likely reflecting localized measurement issues or unresolved sub-grid variability.

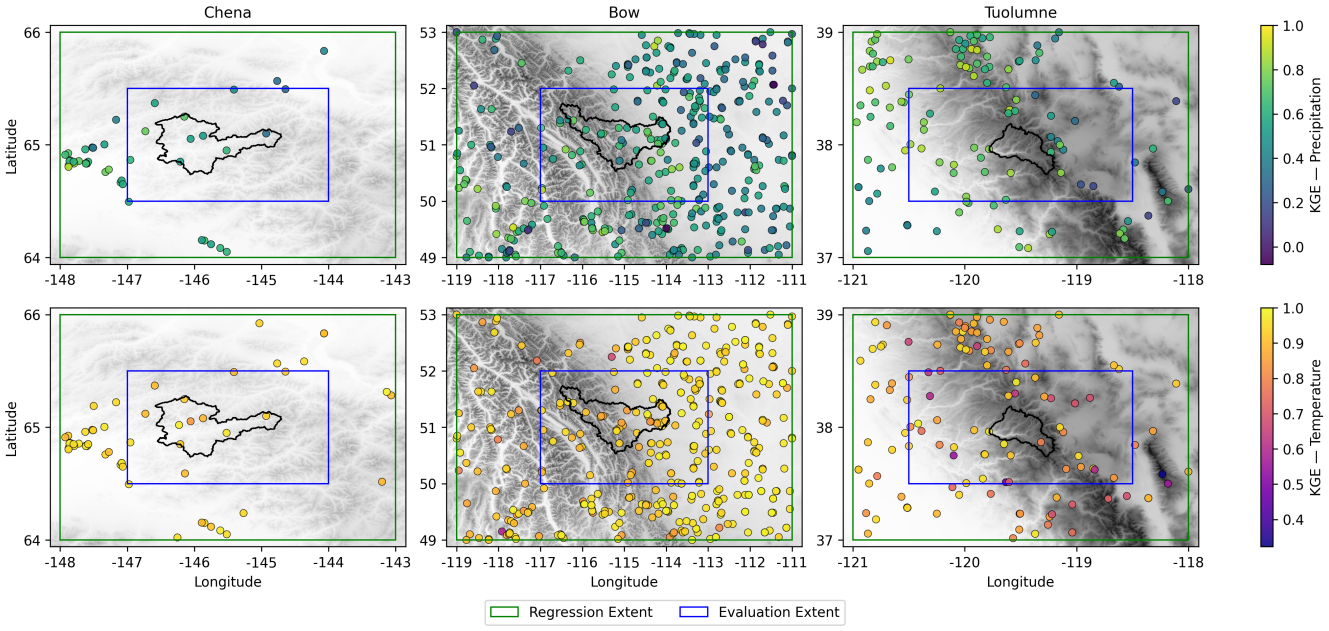


Figure 9. Cross-validation results for Chena, Bow, and Tuolumne River Basins shown as spatial plots for the RF (Random Forest) static and dynamic predictor combinations. Not all stations have both precipitation and temperature measurements. The background map represents elevation.

5.4 Meteorological Ensembles

Using the selected predictor combinations (Table 1), 50-member daily precipitation and temperature ensembles were generated over each study area. To evaluate the probabilistic skill of these ensembles, we computed Brier Skill Scores (BSS) for station-based events, with climatological occurrence as the reference. For precipitation, the evaluated event was daily precipitation at three thresholds (≥ 0.1 mm,≥ 5 mm,≥ 10 mm). For temperature, exceedance events relevant to snow processes were considered (≥ -5 ° C,$T > 0$ ° C, $T > 5$ ° C).

Figures 10 and 11 present the cumulative distribution of BSS across all stations for each basin and configuration, along with non-overlapping block bootstrap uncertainty intervals. Positive BSS values indicate that the ensembles provide more information than climatology, with $BSS = 1$ representing a perfect probabilistic forecast. For precipitation, all RF-based ensembles achieve predominantly positive BSS values, with distributions clearly shifted above zero compared to LWR. Incorporating dynamic predictors consistently increases BSS, demonstrating that reanalysis fields add useful information about event occurrence beyond static topography and climatology. LWR ensembles also show positive skill relative to climatology in most basins but with noticeably lower BSS, reflecting limitations of purely local, linear relationships for capturing event-scale variability. Temperature ensembles exhibit higher and more uniform BSS than precipitation across all methods (Figure 11). RF again outperforms LWR, and static+dynamic configurations yield the highest BSS. The bootstrap uncertainty bands confirm that these differences are statistically robust for the majority of stations. Overall, the ensemble verification indicates that the generated

ensembles are skillful and reliable relative to the climatology. RF regression substantially enhance probabilistic performance, particularly for precipitation. The inclusion of dynamic predictors also provides some minor improvements in all cases.

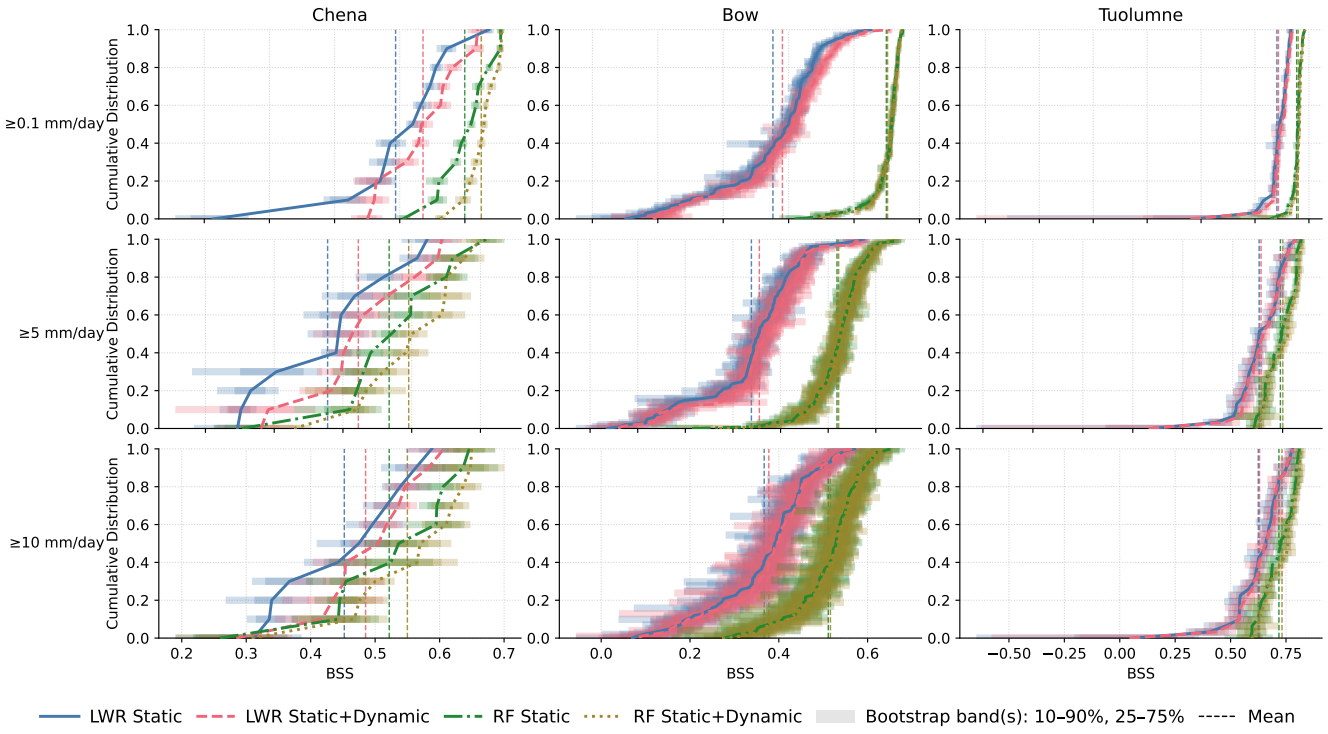


Figure 10. Brier Skill Scores (BSSs) for precipitation with ~~bootstrapping~~bootstrap resampling. The CDFs of BSSs show the relative performance of each method compared to the climatology mean. $BSS = 1$ indicates a perfect probabilistic prediction, while all values above 0 are better than the climatology.

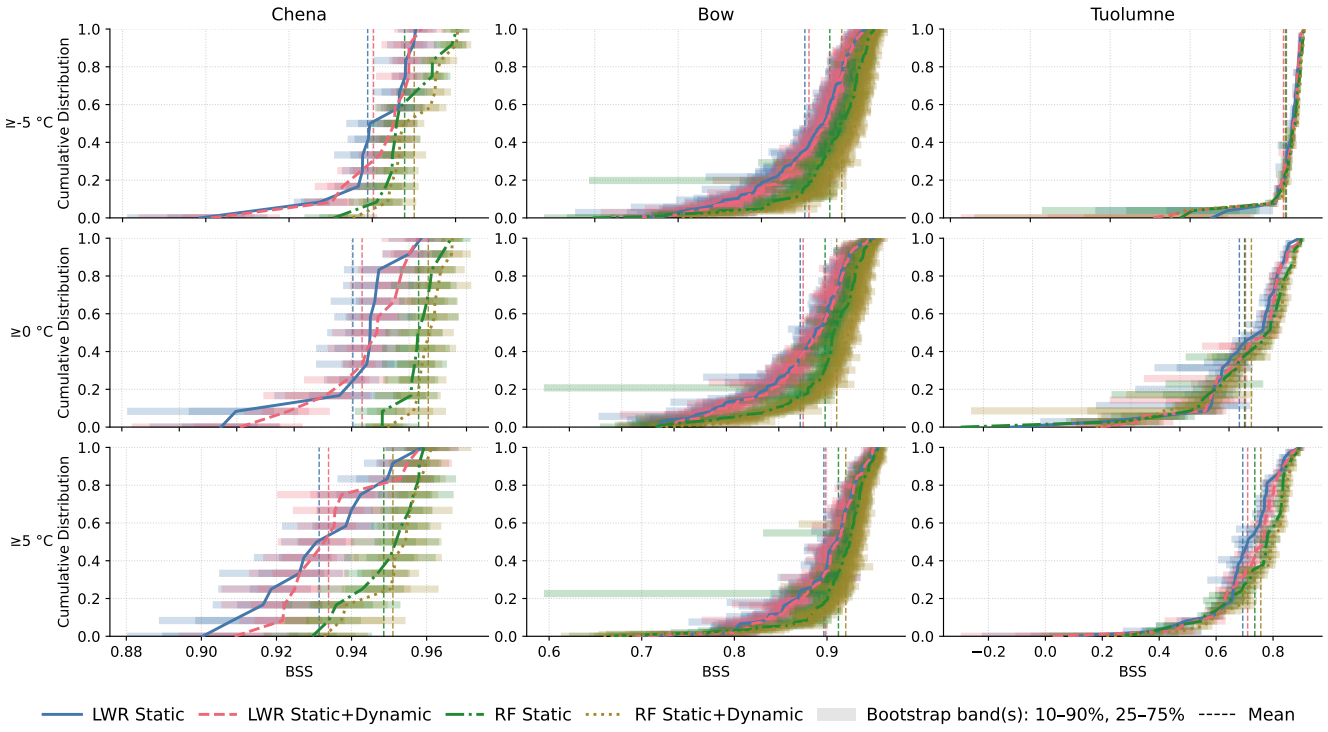


Figure 11. Brier Skill Scores (BSSs) with ~~bootstrapping~~ bootstrap resampling for select temperature thresholds. The CDFs of BSSs show the relative performance of each method compared to the climatology mean. Similar to Figure 10 but for select temperature thresholds.

5.5 Snow Modeling

5.5.1 Evaluation with In situ Measurements

The impact of ensemble meteorological forcing on snow simulations was assessed by driving SUMMA with the 50-member RF ensembles ~~,~~ and two deterministic benchmark forcings (CaSR and ERA5). Simulations were evaluated ~~in~~ for GRUs containing NorSWE manual snow surveys (~~Figure~~Fig. 2), with modeled SWE compared ~~to~~ against observed transect-based ~~SWE~~snow surveys. Figure 12 illustrates SWE time series ~~at~~ for the three GRUs with the ~~highest~~ greatest number of snow surveys in each basin.

The RF-driven ensembles ~~typically exhibit larger SWE magnitudes~~ generally produce larger peak SWE than the deterministic forcings, with observed ~~values lying~~ SWE typically falling within the ensemble spread ~~for most seasons~~during the accumulation season. In the Bow, the ensemble mean ~~SWE is similar in magnitude to CaSR and ERA5, but melt timing is generally better aligned with the observations. In the~~ peak SWE is comparable to the CaSR- and ERA5-driven simulations, whereas in the Chena and Tuolumne ~~, ensemble members capture both~~ the ensembles better capture the range of low- and high-accumulation years, ~~while revealing years where~~ including years when the deterministic forcings substantially under-

or over-estimate peak SWE. Quantitative evaluation of SWE bias and RMSE (Fig. 13) also illustrate that the RF ensembles provide comparable or improved representation of peak SWE magnitude relative to the deterministic forcings.

~~Simulated timeseries evaluated at individual stations can have mixed results. For example, in the Bow station locations where reanalysis CASR and ERA5 simulation values appear to better match SWE observation magnitude than~~ Agreement at individual survey locations is more variable, and in some Bow GRUs the deterministic simulations more closely match the observed SWE magnitude than the RF ensemble mean. ~~This highlights challenges of constraining snow simulations with results that are not completely spatially or temporally representative. However, the overall magnitudes and timing of~~ These differences reflect the challenge of evaluating grid-scale simulations using observations that may not fully represent the spatial variability of snow within a GRU. Consequently, the in situ SWE observations primarily constrain snow accumulation and ~~melt processes do validate the study approach~~peak SWE, whereas snow disappearance date and snow cover duration are evaluated independently using MODIS-derived products (Figure 14). Complete SWE ~~timeseries results, including for snow pillows, are included in Appendix~~time series, including snow pillow evaluations, are provided in Appendix E.

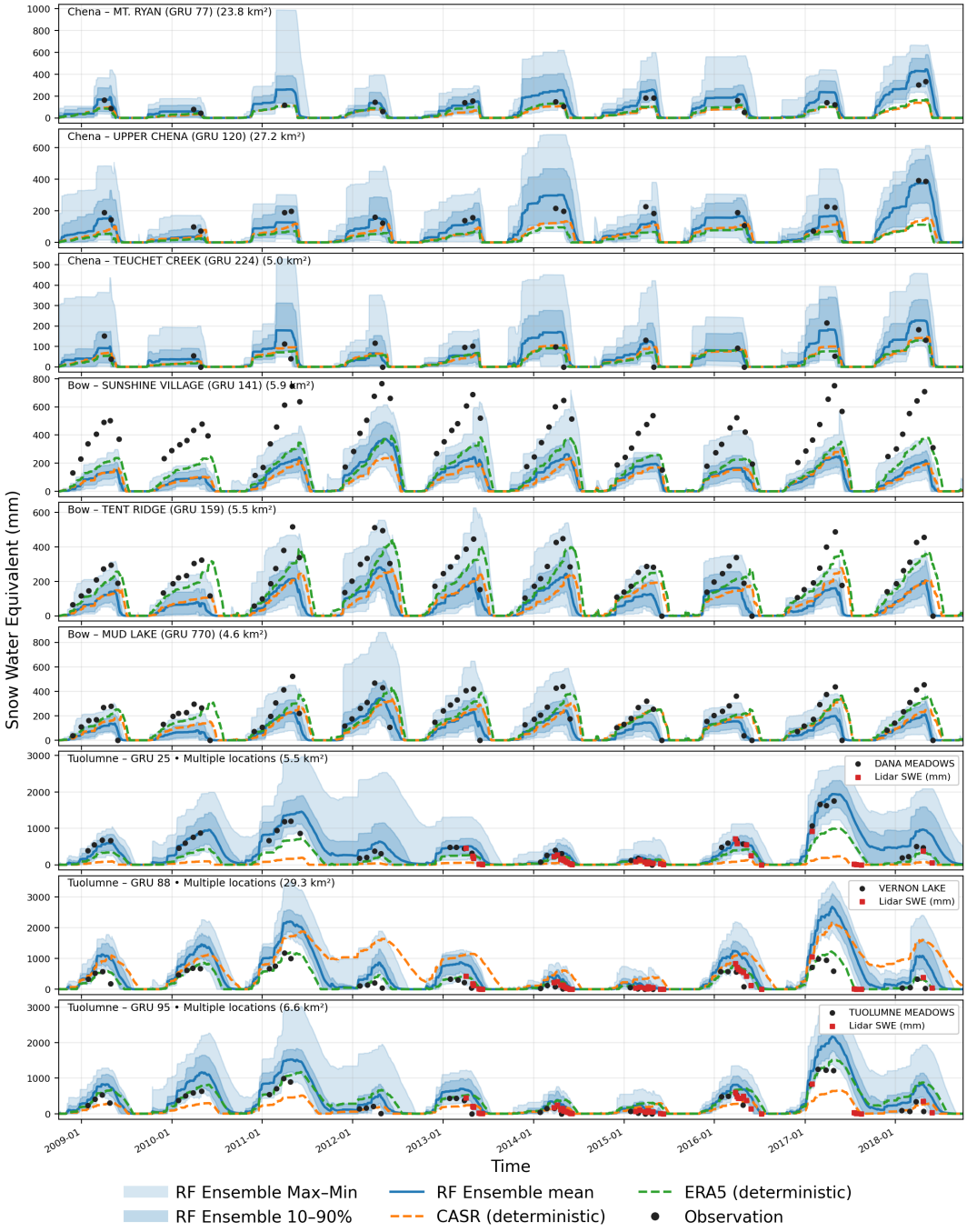


Figure 12. SUMMA sub-basin snow simulations compared to snow surveys. Snow model simulations for the Chena, Bow and ~~Tuolumne Basins~~ Tuolumne are plotted for RF ensembles, CASR, and ERA5 forcing. For each basin the 3 GRUs with the highest amount of snow survey data are shown. For the Tuolumne, lidar SWE estimates averaged over the GRUs are shown.

A summary of SWE performance for each basin is provided by the skill-target diagrams and Taylor diagrams in Figure 13. ~~Each point represents a single GRU–forcing pairing (and ensemble member for the RF ensembles).~~ In the skill-target diagrams, RF ensemble members cluster closer to the origin than ERA5 and CaSR, indicating lower combined pattern and bias error. The ~~spread of RF points reflects the intended representation of forcing uncertainty, but the median skill distance is lower than for the deterministic productsin all three basins. The Taylor diagrams show that the RF ensembles generally achieve correlations with observations comparable to ERA5 and CaSR, while generally having higher values than reference values of normalized standard deviation. There generally is a wider dispersion, which is absent in single-forcing simulations~~RF-based ensemble generally exhibits lower S pattern values than deterministic forcings, which indicates better modeling of the temporal evolution of snow accumulation and melt. This suggests that ensemble forcing better captures the temperature driven processes and event variability that are often smoothed in reanalysis products. In contrast, differences in S bias highlight systematic magnitude errors. Deterministic forcings show more consistent but often biased behavior, with pronounced negative SWE bias in the Bow basin, likely linked to precipitation underestimation and unresolved orographic effects. The ensemble spans both positive and negative bias, reflecting uncertainty in precipitation magnitude and phase partitioning. Outlier GRUs with low skill highlight locations where scale mismatch, local wind redistribution, or other unresolved processes strongly influence SWE, regardless of forcing. Taken together, the snow simulation results demonstrate that the station-based meteorological ensembles translate into physically plausible SWE ensembles in SUMMA, the ensemble spread provides a meaningful envelope that contains the majority of observed SWE values; and the RF-based GPEP ensemble forcings yield equal or better pattern and bias skill than commonly used deterministic reanalyses, while additionally quantifying uncertainty in snow states.

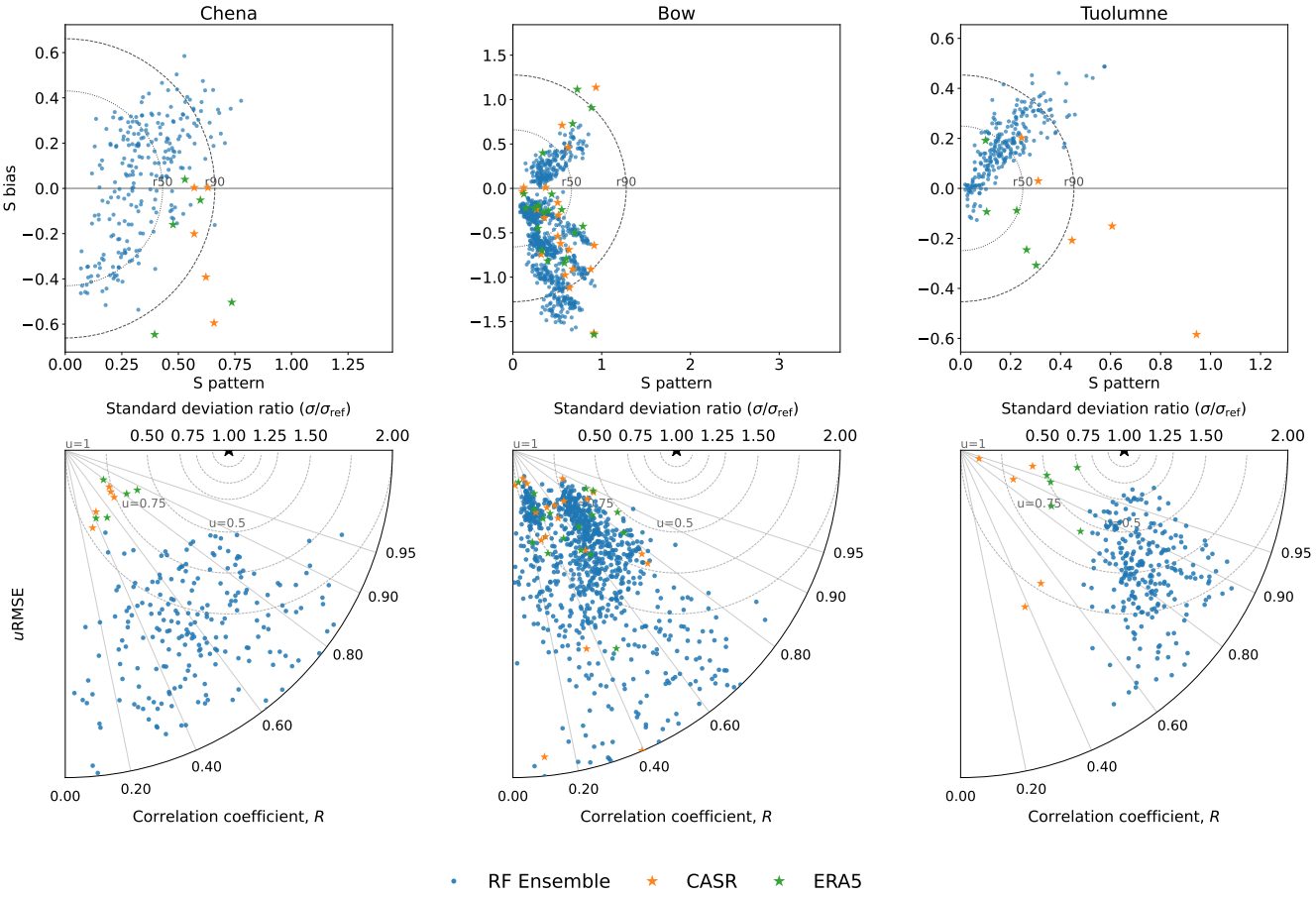


Figure 13. Evaluation of SWE simulations for forcing datasets (Ensemble produced with RF Static–Dynamic predictors (n=50), ~~CASR~~CaSR, and ERA5) using the skill-target (top row) and Taylor diagram (bottom row). Each point represents a single GRU–forcing pairing (and ensemble member for the RF ensembles). The skill-target diagrams (top) display normalized bias (S bias) versus normalized pattern amplitude (S pattern) relative to snow observations; the dashed circles denote 50th and 90th percentiles. Closer to the target origin at zero indicates better performance. The Taylor diagrams (bottom) summarize the same statistics in polar coordinates, with the azimuthal angle representing correlation (R), the radial distance showing normalized standard deviation (S pattern), and contours of equal centered RMSE (u) (Taylor, 2001). Together, the two panels quantify relative skill, bias and correlation consistency following the benchmarking protocol described in Mudryk et al. (2025)

5.5.2 Evaluation with Remote Sensing Snow Metrics

Snow model results can be compared to the MODIS-derived Snow Cover Fraction (SCF) and Snow Disappearance Date (SDD) (Crumley et al., 2020). The results aggregated spatially per water year are shown in Figure 14. A spatial comparison of these results is also informative and provided in Appendix D.

Overall, the modeled snow simulations reproduce the magnitude of both snow cover fraction (SCF) and snow disappearance date (SDD) reasonably well compared with the MODIS-derived estimates. Snow cover duration, represented by SCF, is generally overestimated in the Bow and Tuolumne basins, although the RF ensemble exhibits less bias than the ERA5- and CaSR-driven simulations. Interannual variability in SDD is captured more successfully in the Bow than in the Chena and Tuolumne, where year-to-year variations produced by the model correspond less with the derived measurements.

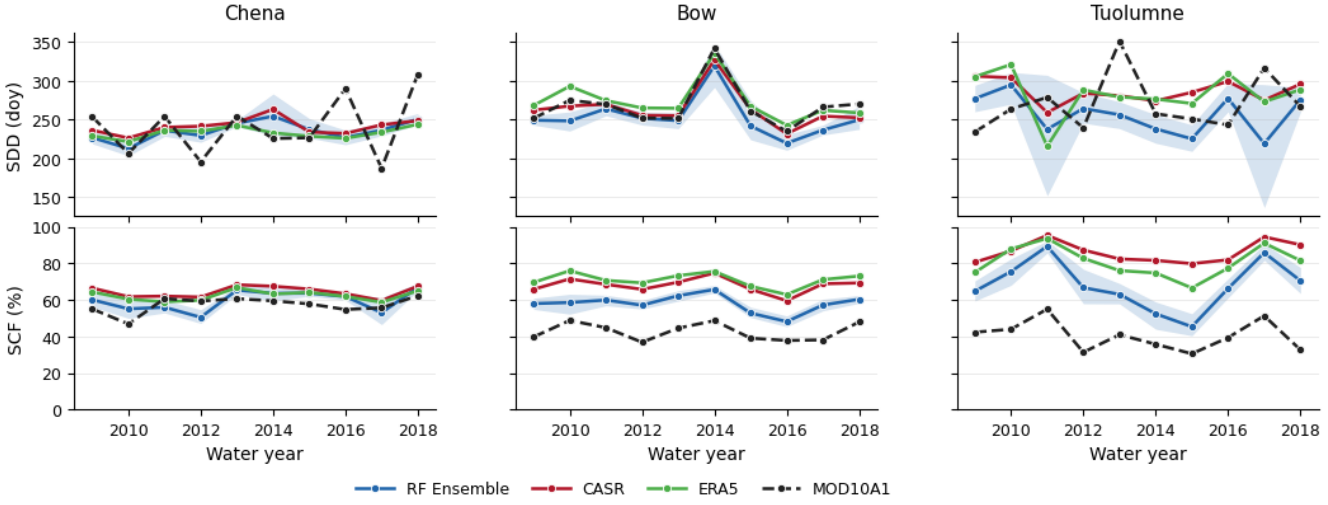


Figure 14. Snow Disappearance Date (SDD) and Snow Cover Fraction (SCF) for snow model results compared to MODIS MOD10A1 derived measurement. Metric are aggregated spatially over the catchments and calculated for each water year

5.5.3 Precipitation Phase Uncertainty

Precipitation phase, expressed here as the fraction of precipitation falling as snow, is shown in as annual climatological cycle in Figure 15. As expected, ensemble uncertainty is greatest during the seasonal transition periods, when precipitation frequently alternates between rain and snow. The Chena exhibits a pronounced peak in ensemble standard deviation during the spring melt period, whereas uncertainty in the Bow is comparatively low. In contrast, the Tuolumne exhibits elevated uncertainty throughout much of the water year, reflecting more persistent uncertainty in precipitation phase. Spatial patterns of precipitation phase, ensemble uncertainty, and normalized uncertainty in water-year snowfall accumulation are presented in Appendix D.

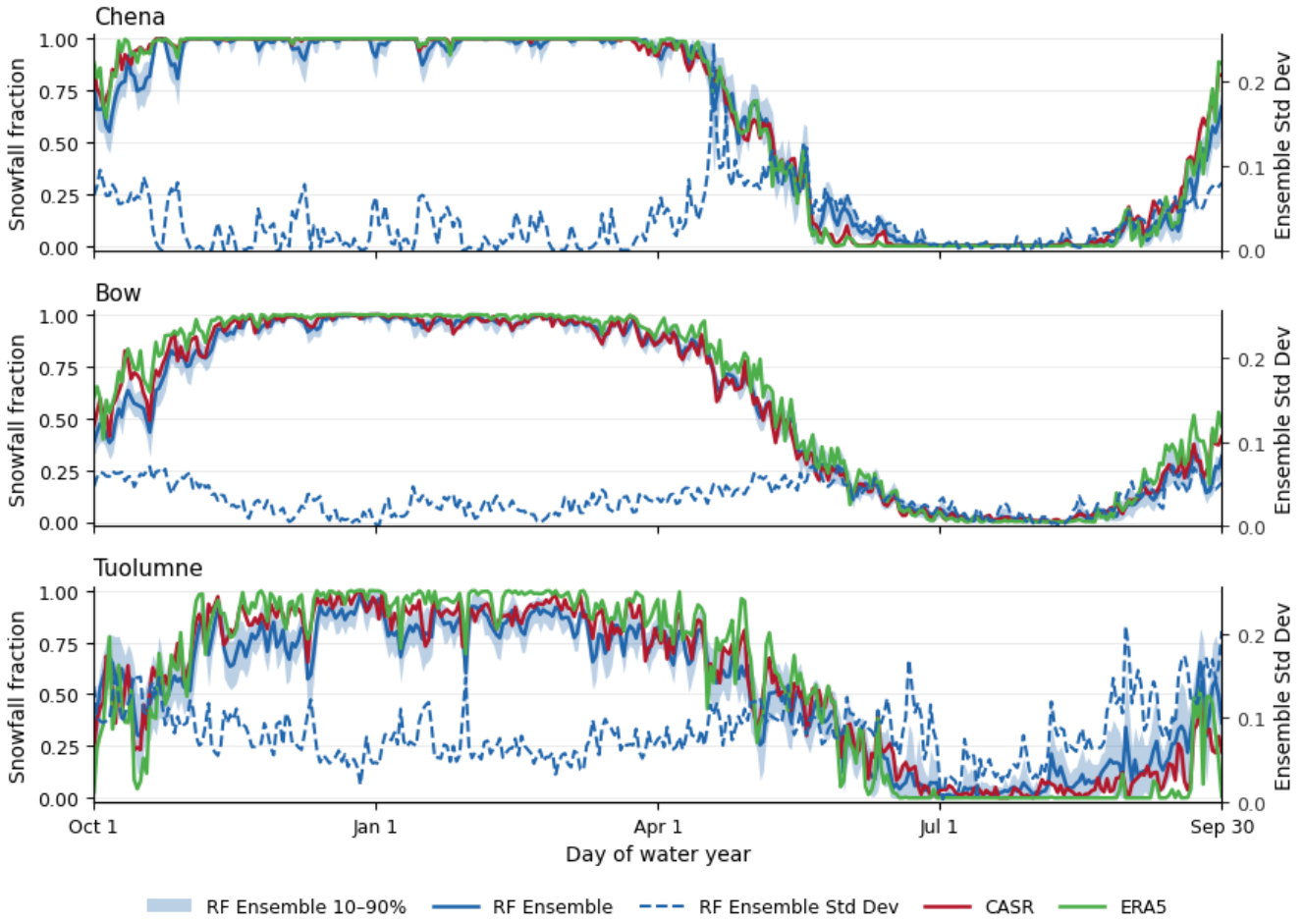


Figure 15. Snowfall fraction (fraction of precipitation that falls as snow) displayed across the water year, including the ensemble spread and standard deviation of ensemble runs.

6 Discussion

This study presents an end-to-end approach to representing meteorological forcing uncertainty in physically based mountain snow simulations. The structured approach integrates undercatch corrected station-based probabilistic meteorology through to an ensemble forcing and distributed energy-balance snow model simulation. Across three contrasting mountain basins, the approach improves deterministic estimation, yields probabilistic forecasts with positive Brier Skill Scores, and produces snow ensembles that capture the magnitude and timing of SWE with credible uncertainty bounds (Figs. 8, 10, 11, 13). The contribution lies not only in the individual components but in their coupling: bias corrected inputs and predictor design improve the estimation of meteorology in complex terrain, and these improvements are realized in physically based modeled snow states.

6.1 Ensemble Mountain Meteorology for Snow Modeling

The results advance a long line of efforts to map meteorology in complex terrain and quantify associated uncertainty. Physiographic approaches such as PRISM demonstrated the value of topographic predictors and expert weighting for gridded climatologies (Daly et al., 2008, 2021), while operational products such as seNorge and related systems have emphasized deterministic spatial continuity, but rarely provide event-scale uncertainty (Lussana et al., 2018). Probabilistic geostatistical frameworks, including GMET and subsequent extensions, established that regression against static predictors combined with stochastic sampling of spatially correlated residuals can generate calibrated ensembles (Clark and Slater, 2006; Newman et al., 2015). Our work builds directly on that foundation in two ways: it systematically incorporates dynamic predictors from reanalyses alongside static topography, and it evaluates machine-learning estimators against locally weighted regression within the same ensemble architecture. The finding that RF outperform LWR for precipitation and temperature in multiple mountain settings strengthens earlier indications that non-linear learners can better represent interactions between orography and time varying patterns (Wolfensberger et al., 2021; Baez-Villanueva et al., 2020).

The Large Sample Emulator (LSE) provides a scalable approach for calibrating physically based snow models across large domains. This build on recent work by Tang et al. (2025b) and (Farahani et al., 2025) in application in large domain modeling. Rather than relying on basin-specific manual calibration, the emulator transfers information from a diverse set of snow monitoring sites to estimate locally appropriate parameter sets based on climatic and physiographic attributes. This enables spatially varying parameterization while maintaining a consistent calibration methodology across the three study basins. Although the LSE does not reduce structural uncertainty within SUMMA, it provides a reproducible framework for parameter estimation that complements the explicit treatment of meteorological forcing uncertainty presented in this study.

A further contribution is the explicit propagation of this quantified forcing uncertainty into a physically based, energy-balance snow model and its evaluation against ~~in-situ~~ in situ SWE observations. Mountain snow modeling studies, whether in multi-model intercomparisons or basin-specific analyses, most often rely on a single deterministic forcing or meteorologic "ensembles of opportunity" constructed from independent reanalyses (Essery et al., 2013; Henn et al., 2018b; Schreiner-McGraw and Ajami, 2022). While useful for assessing dataset spread, such ensembles do not reflect the underlying interpolation and representativeness uncertainties tied to the observing network. In contrast, the ensembles developed here are explicitly conditioned on stations and incorporate spatially structured residual variability, establishing a clearer link between observational density, predictor choice, and modeled uncertainty. The improved representation of SWE magnitude and melt timing relative to ERA5 and CaSR demonstrates the value of carrying realistic input uncertainty through an energy-balance model. The inclusion of MODIS-derived Snow Cover Fraction (SCF) and Snow Disappearance Date (SDD) complements the evaluation based on in situ SWE observations by providing basin-wide assessment of snow persistence and melt timing. While snow surveys constrain SWE magnitude at discrete locations, remotely sensed metrics evaluate the spatial consistency of simulated snow evolution across each watershed. The generally good agreement between modeled and MODIS-derived snow metrics, though clear biases exist in SCF due to spatial patterns that vary with topography.

The snow simulation results show that the ensemble framework improves representation of snowpack temporal dynamics while explicitly capturing uncertainty in forcing. Rather than uniformly outperforming deterministic products, the ensemble provides a physically meaningful range of outcomes that reflects sensitivity to precipitation magnitude, phase, and variability. This reveals both improved and degraded simulations across sites, highlighting where deterministic forcings are biased and where snow processes are most sensitive to meteorological uncertainty.

6.2 Methodological Implications

~~The behavior of the~~

In selecting predictors and estimators ~~is~~ for meteorological ensembles, the findings of this study are consistent with process understanding and with lessons from physiographic mapping. Elevation remains the dominant control on temperature, and moderate spatial smoothing of terrain derivatives improves estimation for both variables by emphasizing synoptic structure while retaining key orographic gradients. Dynamic predictors add physically relevant variability that static physiography cannot supply, aligning ~~the estimation problem~~ with the drivers used ~~by land~~forland-surface models. The RF advantage over LWR suggests that non-linear interactions among static and dynamic fields are important in sparse mountain networks; at the same time, the LWR framework retains value where transparency and locality are priorities. Because ensemble reliability depends on the fidelity of both the mean and the residual fields, the deterministic gains translate directly into higher Brier Skill Scores once residuals are sampled with spatial correlation.

Workflows such as the one developed for this paper ~~(which used a package called Snakemake)~~ using Snakemake can also contribute to reproducibility and transferability (Köster and Rahmann, 2012). Implementing GPEP with rule-based automation can facilitate consistent application across regions, straightforward selection of predictor sets, and verification. In this study, the use of a standardized discretization (TDX-Hydro) and a consistent SUMMA configuration, coupled with emulator-assisted parameter selection, eased comparative inference across basins. Although such an end-to-end workflow adds an additional software wrapper, as well as dependencies and constraints in conducting such analyses, it also can provide efficiencies in generalizing the analysis~~, in contrast to methods used in local studies that may rely on tailored or manual data and preprocessing steps that are time-intensive to apply to new locations~~.

6.3 Potential Application and Limitations

The ~~results indicate~~ study approach provides a practical route for operational agencies and research groups that require uncertainty-aware snow estimates. The approach can harness region specific station measurements and recent gridded analysis to produce skillful ensemble forcing. These data are available in real time, and the procedure can be automated. Driving an energy-balance snow model with these ensembles yields SWE state estimates with credible spread and timing that are directly usable in data assimilation, seasonal outlooks, and risk analyses. The demonstration across three distinct basins suggests that the approach is robust to differences in observing density and topographic complexity, which is essential for scaling to new regions.

Limitations remain across many stages of this approach. Undercatch correction relies on simplifying assumptions on gauge type that should be refined where suitable gauge metadata is available (Huidobro et al., 2023). While reanalysis winds provide a reasonable surrogate for undercatch correction, developing downscaled, gauge-height wind fields for mountains remains a clear priority (Mott et al., 2018; Berg et al., 2024; Marsh et al., 2023; Helbig et al., 2017). The absence of consistent gauge metadata in global archives continues to be a major source of uncertainty for undercatch corrections, reinforcing the need for curated, operator-specific metadata where available (Mekis et al., 2018; Smith et al., 2022; Sun et al., 2019). Daily ensembles subsequently disaggregated into sub-daily forcings limit the representation of diurnal energy fluxes and short-duration events; extending GPEP to hourly resolution could improve melt onset and rain-on-snow timing. The correlation structure and transformation chosen for precipitation influence extremes and dispersion and merit a targeted study. Also, the SUMMA configuration used here does not explicitly represent wind redistribution or avalanching, and GRU-scale evaluation against point or transect observations entails scale mismatch. ~~This point is illustrated in Figure 16, where measurements collected in a well-instrumented research catchment with different localized settings within a single GRU yield highly different SWE magnitudes.~~ Furthermore, a single model physics and parameter set is used in this study, so exploration of multi-physics and parameter ensembles could improve probabilistic representation. Finally, further benchmarking with spatial products, such as airborne lidar, could further disentangle and quantify forcing and structural uncertainties.

A key limitation of this framework is the representation of snow processes at the GRU scale, which aggregates substantial sub-grid heterogeneity. Figure 16 illustrates this limitation, showing that SWE measurements collected under contrasting local conditions within a single GRU can differ markedly. Preferential accumulation areas, including wind-sheltered slopes, leeward deposition zones, and forest openings, often sustain deeper and more persistent snowpacks than surrounding terrain, yet are not explicitly resolved. These locations are also underrepresented in meteorological and snow observation networks, compounding uncertainty in both forcing and model evaluation particularly in melt periods. Although the ensemble framework partially accounts for unresolved variability through spatially correlated perturbations, it does not explicitly represent slope-scale processes such as wind redistribution and preferential deposition. Addressing these limitations will require finer spatial discretization and higher-resolution meteorological forcing, with emerging regional reanalyses and atmospheric models offering improved representation of orographic precipitation and other terrain-driven processes.

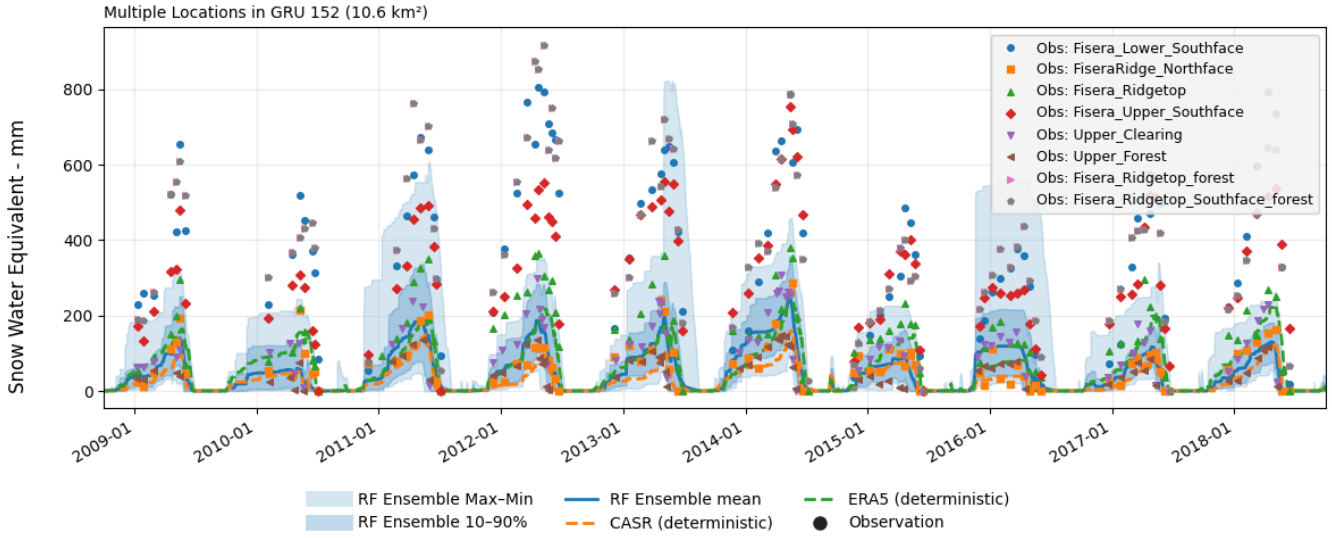


Figure 16. SUMMA SWE simulations of a single GRU with multiple observation sites in the Bow. Measurement location has a large effect on the quantity of snow. This highlights the challenge of evaluation and interpretation of semi-distributed model simulations with ~~in-situ~~ in situ data.

In summary, this work extends probabilistic geospatial estimation from the station network to physically based snow modeling and demonstrates that doing so improves both deterministic skill and probabilistic reliability in complex terrain. Relative to the existing literature, it integrates undercatch correction within ensemble construction across multiple catchments, leverages dynamic predictors with a non-linear estimator to enhance station-based ensembles, and shows that the resulting forcings translate into more credible SWE simulations than commonly used deterministic products. The combination of methodological transparency, open implementation, and multi-basin evaluation provides a clear pathway to improve probabilistic snow modeling and water prediction.

7 Conclusions

This study presents a structured approach for propagating meteorological uncertainty through physically based mountain snow modeling. The workflow begins with bias correction of precipitation observations, followed by geospatial estimation of meteorological variables and the generation of high-resolution ensemble meteorological fields. These ensembles capture the spatial and temporal variability of forcing data in complex terrain and are used to drive physically based snow simulations, which propagate input uncertainty into snow estimates. The ~~approach is applied~~ snow model parameters were calibrated using a Large Sample Emulator. The overall approach was applied and assessment in three contrasting mountain basins: the Chena in Interior Alaska, the Bow in the Canadian Rockies, and the Tuolumne in the Sierra Nevada. These catchments represent a range of climatic conditions, monitoring contexts, and ~~density of meteorological station observations~~meteorological station

densities. Applying this framework within the SUMMA modeling system enabled consistent and transparent representation of meteorological uncertainty in snow state predictions across these diverse settings.

Addressing key deficiencies in mountainous meteorology estimation, namely the influence of complex topography and ~~wind induced~~ wind-induced gauge undercatch, ~~helps to improve~~ improves meteorological estimates. Applying WMO-standard undercatch corrections improves the magnitude of cold-season precipitation estimates. Reanalysis wind data provide a practical supplement where in situ wind measurements are unavailable, enabling consistent correction across large domains. Topographic descriptors, particularly elevation and smoothed terrain derivatives, remain key predictors of both precipitation and temperature patterns. Incorporating dynamic atmospheric predictors from reanalysis further enhances model performance, particularly for precipitation, by capturing synoptic-scale variability that topography alone cannot explain.

Among the deterministic estimation methods, RF-based meteorological estimation ~~models~~ consistently outperformed locally weighted regression (the prior default method in studies applying GMET and GPEP), achieving higher cross-validation skill and greater probabilistic reliability once ensemble dressing with spatially correlated errors was applied. Ensemble meteorological forcing fields generated through this framework exhibited markedly higher Brier Skill Scores than climatological references, demonstrating that the ensembles are both reliable and informative. When used to drive physically based snow simulations, these ensembles improved the representation of uncertainty in snow water equivalent, providing a more credible basis for assessing snow-driven hydrologic variability.

Overall, this work highlights the importance and opportunity of explicitly representing meteorological uncertainty in mountain snow modeling. Probabilistic meteorological forcing estimation, when coupled with physically based modeling, provides an effective strategy to bridge the gap between sparse observations and complex terrain processes. The framework nevertheless remains subject to uncertainties arising from precipitation undercatch correction, GRU-scale representation of heterogeneous snow processes, and the use of a single model structure and parameterization, highlighting opportunities for future development. Future developments should aim to extend this ensemble methodology to sub-daily timescales~~and~~, improve representation of sub-grid mountain processes, incorporate multi-physics and parameter ensembles, and integrate it with data assimilation systems to dynamically constrain snowpack states. As mountain hydrology faces growing challenges from climate variability and changing water demands, probabilistic, physically based approaches such as this offer a foundation for ~~both~~ scientific investigation and operational prediction that explicitly include uncertainty estimates.

Code and data availability.

All data and code used in this study are openly available. The primary station-based meteorological data are the Station-Based Serially Complete Earth Dataset (SCDNA; Tang et al., 2021), available through the US National Center for Atmospheric Research (NCAR) data repository. Topographic descriptors were derived from the MERIT Hydro digital elevation dataset (Yamazaki et al., 2017), which provides hydrologically conditioned elevation and drainage information at 3-arcsecond resolution. Atmospheric reanalysis predictors were obtained from the European Centre for Medium-Range Weather Forecasts (ERA5;

Hersbach et al., 2020) and the Canadian Surface Reanalysis (CaSR v3.1; Gasset et al., 2021), both publicly accessible through their respective data portals.

All code for data processing and model coupling is open source and available via GitHub. The workflows for data preprocessing, undercatch correction, and geospatial probabilistic estimation are implemented in the repository `gpep_snakemake` (Casson, 2025b). The workflow used to couple ensemble meteorological forcings to the SUMMA model framework is available in `gpep_to_summa_snakemake` (Casson, 2025c). Both repositories include reproducible Snakemake pipelines, configuration files, and documentation for applying the Geospatial Probabilistic Estimation Package (GPEP; Tang et al., 2024) and the SUMMA hydrological model across diverse mountain basins.

The ensemble meteorological datasets and models generated for all three study basins, the Chena, Bow, and Tuolumne, along with the corresponding processed station and topographic data used in this study, are available via Zenodo and the Federated Research Data Repository (FRDR) (Casson, 2025d, a). These archives include ensemble forcing files, SUMMA model files, GPEP input files and basin-level metadata to support reproducibility and further research on probabilistic snow modeling in mountainous environments.

Appendix A: Undercatch Correction Methodology

Detailed equations, constants, and assumptions for the undercatch correction methods applied in this study are provided in Table A1. The WMO SPICE UTF (Kochendorfer et al., 2017a, b) is applied in the Bow Basin, while the WMO (1998) standard rain gauge correction (Goodison et al., 1998; Yang et al., 1998) is used for the Chena and Tuolumne basins. Wetting loss are not included in correction. Each approach assumes consistent gauge type and shielding within each basin, reflecting available metadata and historical network standards.

Method	WMO SPICE Universal Transfer Function (2017)	WMO Solid Precipitation Measurement Intercomparison (1998)
General Information		
Key Reference	Kochendorfer et al. (2017a, b)	Goodison et al. (1998)
Study Areas	Bow (Canada)	Chena and Tuolumne (United States)
Assumed Gauge Type	Geonor and Pluvio	US standard 8" gauge
Shielding	No shield	No shield
Input Data		
SC-Earth Station Data	P: precipitation (uncorrected, mm/day); U: wind speed (m/s); T: air temperature (°C)	P: precipitation (uncorrected, mm/day); U: wind speed (m/s); T: air temperature (°C)
ERA5 Reanalysis	U: daily max wind speed (m/s), daily mean air temperature (°C)	U: daily max wind speed (m/s), daily mean air temperature (°C)
Correction Equations		
Catch Efficiency	$CE = e^{-aU}\left[1 - \tan^{-1}(bT) + c\right]$	$CE = \exp\left(d - eU^{f}\right)$
Precipitation Partitioning	Implicit in correction formula	$S = 1 - \frac{1}{1 + e^{a_0 + a_1 T}}$
Threshold Behaviour	$U < U_{\text{thres}};\ T > T_{\text{thres}},\ CE = 1$	$U < U_{\text{thres}}$
Wind Speed Correction	Transfer Function for 10m available	Logarithmic wind profile correction
Precipitation Correction		
Corrected Precipitation	$P_{\text{corr}} = \frac{P}{CE}$	$P_{\text{corr}} = (1 - S)P + \frac{S}{CE}P$
Additional Information		
Additional References	Smith et al. (2022)	Newman et al. (2020); Froidurot et al. (2014); Huidobro et al. (2023)

Table A1. WMO gauge undercatch methodologies used to correct precipitation measurements in Canadian and US catchments. A simplifying assumption of consistent gauge type within each study area is applied due to limited metadata on gauge configuration.

Undercatch correction implementation was compared to the benchmark dataset of Smith et al. (2022) who implemented SPICE UTF methods for stations across Canada. This was performed for overlapping stations (n=10) surrounding the Bow watershed. This comparison highlights differences between shielding assumption configurations, the use of ERA5 wind and correcting a comparison of 2 m compared to 10 m measurement heights. This comparison was performed for the Bow watershed as the assumption of consistent gauge assumptions in United States gauging networks is consistent with (Newman et al., 2020; Huidobro et al., 2023). As depicted in Figure A1, the methodology applied to SC-Earth data matches the seasonality and magnitude of the benchmark dataset. The use of ERA5 winds leads to conservative undercatch compared to ~~in-situ~~ in situ data. The selection of shielding configuration is a key assumption, highlighting the need for station metadata across gauging networks.

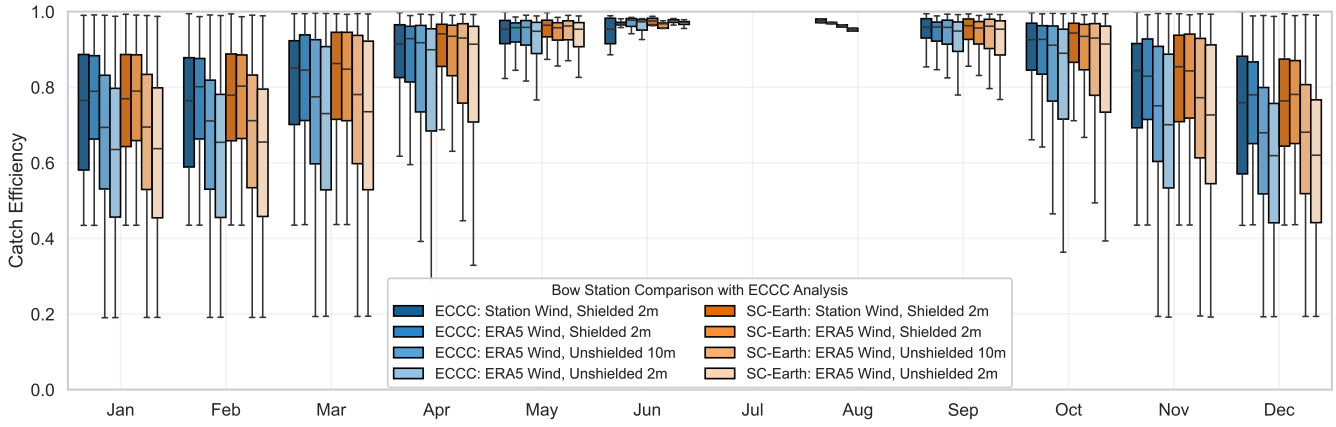


Figure A1. Comparison of methodological choices for undercatch correction using overlapping SC-Earth stations with that of Smith et al. (2022) in the Bow. The ECCC station dataset, shielded 2 m represents the reference dataset, while separate iterations compare supplementing SC-Earth station data, ERA5 wind and the change with wind height corrections.

Appendix B: GPEP Methodology

This appendix provides the statistical methods used in the Geospatial Probabilistic Estimation Package (GPEP) for this study. A more complete description of the software and computational considerations is available in Tang et al. (2024). A synthesis of the GMET method development prior to the Python GPEP implementation is available also available in Appendix B of Bunn et al. (2022).

B1 Deterministic Estimation

Deterministic estimation forms the first step of the probabilistic workflow. This step estimates the gridded mean of the target variable (e.g., precipitation or temperature measured at station locations) and for intermittent variables the probability of occurrence. These deterministic fields, calculated with ~~Locally Weight Regression and Random Forest~~locally weighted regression and random forest, form the basis for the ensemble generation process.

B1.1 Locally Weighted Regression

Locally weighted regression (LWR) is used to generate deterministic grid estimate. For each grid cell, the relationship between the observed or transformed variable x_o and a set of predictor variables A_i is described by

$$x_o = \hat{x} + \epsilon = \beta_0 + \sum_{i=1}^{n} A_i \beta_i + \epsilon, \tag{B1}$$

where A_i is the ith predictor, β_0 and β_i are regression coefficients, and ϵ is the residual error term. Spatial localization is achieved using distance-based weights that limit the influence of observations to those stations most relevant to the grid

cell. This approach allows the regression to capture regional-scale variations while maintaining computational efficiency. As demonstrated by Bunn et al. (2022), the same formulation can incorporate temporally varying predictors such as NWP or reanalysis data.

For variables that are intermittent, such as precipitation, the probability of occurrence is estimated using a locally weighted logistic regression. This calculates the Probability of Precipitation (POP), defined as

$$\text{POP} = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{i=1}^{n} A_i \beta_i\right)\right)}. \tag{B2}$$

The resulting mean field and probability field are then used in the probabilistic estimation step to generate ensemble realizations that reflect both the expected value and its associated uncertainty.

B1.2 Random Forest Regression and Classification

Random ~~Forest~~ forest (RF) methods provide a non-parametric alternative to locally weighted regression for generating the deterministic estimate. RF models can be applied in both regression and classification modes, corresponding respectively to the estimation of continuous variables (e.g., precipitation amount, temperature) and the prediction of event probabilities (e.g., precipitation occurrence) (Wolfensberger et al., 2021). Within the geospatial probabilistic estimation framework, RF regression and classification serve analogous roles to the locally weighted linear and logistic regression, respectively.

An RF model consists of an ensemble of decision trees, each trained on a random subset of the available data and predictor variables (Breiman, 2001). This process introduces stochasticity in both the sampling of training observations (via bootstrapping) and the selection of predictors at each node, ensuring that the ensemble captures a diverse range of predictor–response relationships. The final prediction is obtained by aggregating the outputs of all trees, which improves generalization and reduces overfitting relative to a single decision tree (Cutler et al., 2007).

In regression applications, each tree produces an individual prediction $\hat{x}_j$ of the variable of interest x_o, based on a set of predictors $A_1, A_2, \ldots, A_n$. The ensemble mean prediction $\hat{x}$ is then computed as the arithmetic average across all T trees:

$$\hat{x} = \frac{1}{T} \sum_{j=1}^{T} \hat{x}_j, \tag{B3}$$

where T is the number of trees in the forest and $\hat{x}_j$ is the prediction from the jth tree. Each decision tree recursively partitions the predictor space by selecting split thresholds that minimize a specified cost function. The averaging of multiple decorrelated trees yields a stable estimate that reflects nonlinear relationships between predictors and responses without requiring an explicit parametric model (Tyralis et al., 2019).

In classification applications, each tree predicts a categorical outcome $\hat{y}_j$, such as the presence or absence of precipitation. The ensemble classification $\hat{y}$ is determined by majority voting across the T trees:

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T), \tag{B4}$$

where mode$(\cdot)$ denotes the most frequent class among the tree-level predictions. The result is a probability of occurrence derived from the fraction of trees that predict a given outcome, which can be used as the Probability of Precipitation (POP) in probabilistic estimation.

The RF models in GPEP are implemented using the scikit-learn library, enabling straightforward implementation (Pedregosa et al., 2011). Unlike locally weighted regression, RF models are inherently global: they are trained once using all available stations and predictor data, rather than fitting a separate model at each grid cell. This global structure allows the RF approach to exploit complex, nonlinear dependencies between predictors (e.g., topography, dynamic atmospheric predictors, and spatial coordinates), while maintaining computational scalability across large domains (Fox et al., 2020).

Random ~~Forests~~forests, therefore, provide a robust and flexible alternative for deterministic gridded estimation of geospatial variables, particularly in regions where predictor–response relationships are highly nonlinear or where the spatial sampling of observations is sparse or irregular (Sekulić et al., 2020). The mean field and probability field outputs from RF regression and classification are subsequently used in the probabilistic estimation step, where uncertainty is represented through spatially correlated ensemble generation.

B2 Probabilistic Estimation and Ensemble Generation

The probabilistic estimation stage converts the deterministic fields into ensemble realizations that represent the spatial and temporal uncertainty of the estimated variables. This is achieved by combining the deterministic mean estimates, the associated uncertainty calculated from station error, and a stochastic component represented by spatially correlated random fields (SCRFs). The SCRFs ensure that the generated ensembles preserve realistic spatial coherence and correlation structures that are consistent with those observed in the station data and predictor fields.

In the general case, a variable X at a given grid cell and time step is represented as a random variable drawn from a Gaussian probability distribution, written in-text as $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ corresponds to the deterministic mean estimate and σ quantifies local uncertainty. For variables such as temperature and temperature range, this normal approximation is appropriate. For intermittent and skewed variables such as precipitation, the distribution is decomposed into two components: the probability of occurrence (p_0) and the conditional distribution of nonzero values. The conditional precipitation field is transformed to normal space using a Box–Cox or power-law transformation to ensure approximate Gaussianity (Newman et al., 2015; Tang et al., 2024).

To represent spatial structure, ensemble perturbations are generated using SCRFs sampled from a standard normal distribution. The SCRFs impose correlation among neighboring grid cells using an exponential correlation model:

$$c_{i,j} = \exp\left(-\frac{d_{i,j}}{L_c}\right), \tag{B5}$$

where $c_{i,j}$ is the correlation between grid cells i and j, $d_{i,j}$ is the distance between them, and L_c is the spatial correlation length. The correlation length is estimated empirically from station data and varies by variable and domain. SCRFs are gen-

erated using a conditional simulation approach (Clark and Slater, 2006), ensuring that each random field realization maintains both the specified correlation structure and the correct marginal statistics.

Temporal coherence is introduced by linking random fields across successive time steps using lag-1 autocorrelation coefficients derived from the observed data. Cross-variable correlations (for example, between precipitation and temperature range) can also be imposed to maintain realistic multivariate relationships (Tang et al., 2024).

For each ensemble member, the SCRFs are used to perturb the deterministic estimates according to their local uncertainty. For normally distributed variables such as temperature, the probabilistic estimate is computed as

$$x = \mu + R \cdot \sigma, \tag{B6}$$

where R is the corresponding value from the SCRF in that grid cell and time. For intermittent variables such as precipitation, ensemble generation proceeds in two steps. First, the probability of occurrence is evaluated: if the cumulative probability of R is less than the occurrence probability p_0, the grid cell is assigned zero precipitation. Otherwise, the nonzero precipitation amount is computed by scaling and back-transforming the random draw in normal space:

$$y = \begin{cases} 0, & \text{if } F_N(R) \le p_0, \\ \mu + F_N^{-1}\left(\frac{F_N(R)-p_0}{1-p_0}\right)\sigma, & \text{if } F_N(R) > p_0, \end{cases} \tag{B7}$$

where $F_N(\cdot)$ and $F_N^{-1}(\cdot)$ are the standard normal cumulative distribution and its inverse. The resulting y values are then inverse-transformed from normal space to yield precipitation in physical units. This process is repeated independently for each ensemble member, generating a suite of spatially and temporally coherent realizations.

Appendix C: ~~Surface Meteorology Inputs~~ Large Sample Emulator Errors Compared to SUMMA

~~Variable Description Units CASR v3.1 variable ERA5 variable time Time s time time precip Precipitation rate kg m^{-2} s^{-1} A_PR0_SFC pptrate airtemp Air temperature K P_TT_1.5m airtempSWRadAtm Downward shortwave radiation W m^{-2} P_FI_SFC SWRadAtm LWRadAtm Downward longwave radiation W m^{-2} P_FB_SFC LWRadArm windspd Wind speed m s^{-1} P_UVC_10m windspd airpres Air pressure Pa P_P0_SFC airpres spechum Specific humidity kg kg^{-1} P_HU_1.5m spechum Summary of surface meteorology data input requirements for the SUMMA energy balance snow model, with corresponding CASR v3.1 and ERA5 variable names.~~

Appendix D: ~~Model Calibration Parameter Ranges~~

~~Parameter name in SUMMA Lower Upper Transformation Units tempCritRain 272.16 274.16 KtempRangeTimestep 0.5~~ For the 162 stations used to train the emulator (in red in Figure C1), we compared the Taylor Skill Score (TSS; Taylor, 2001)

from SUMMA (i.e., simulations vs. observations) and the emulated TSS value using random forecast (Figure C1b). The emulator comprises 162,000 SUMMA simulations (162 stations × 1,000 parameter sets; density dots in Figure C1b), for step $i = 5$ of the iterative process. To obtain new parameter sets for each station, we search the emulated response surface using a Genetic Algorithm (GA), yielding 100 new parameter sets per station. SUMMA is then run for these parameter sets to compute TSS, providing an evaluation error for parameter sets unseen during random forecast training. At step i = 5~~KfrozenPrecipMultip 0.95 1.05 albedoMax 0.7 0.95 albedoMinWinter 0.4 0.6 albedoMinSpring 0.35 0.5 albedoDecayRate 100000 5000000 albedoRefresh~~ , the MAE between emulated and SUMMA TSS values for these unseen parameter sets is 0.242 TSS units. We also apply the trained emulator from each iteration step to search for parameter sets at unseen stations (shown in red in Figure C1c, n = 608), again using a GA to produce 100 candidate parameter sets per station. The error between emulated and SUMMA TSS values is then computed across these 608 stations × 100 parameter sets. As shown in Figure C1d, this error decreases steadily across iteration steps, reflecting both the growing training dataset — which increases from 500 to 1~~10 Fcapil 0.01 0.1 mw_exp 1~~ ,000 parameter sets per station between i = 0 and i = 5 ~~z0Snow 0.0001 0.03 log10m SUMMA snow parameters used in the Large Sample Emulator (LSE) calibration approach as well as the lower and upper bounds for parameter search.~~ — and the progressively better-explored parameter space.

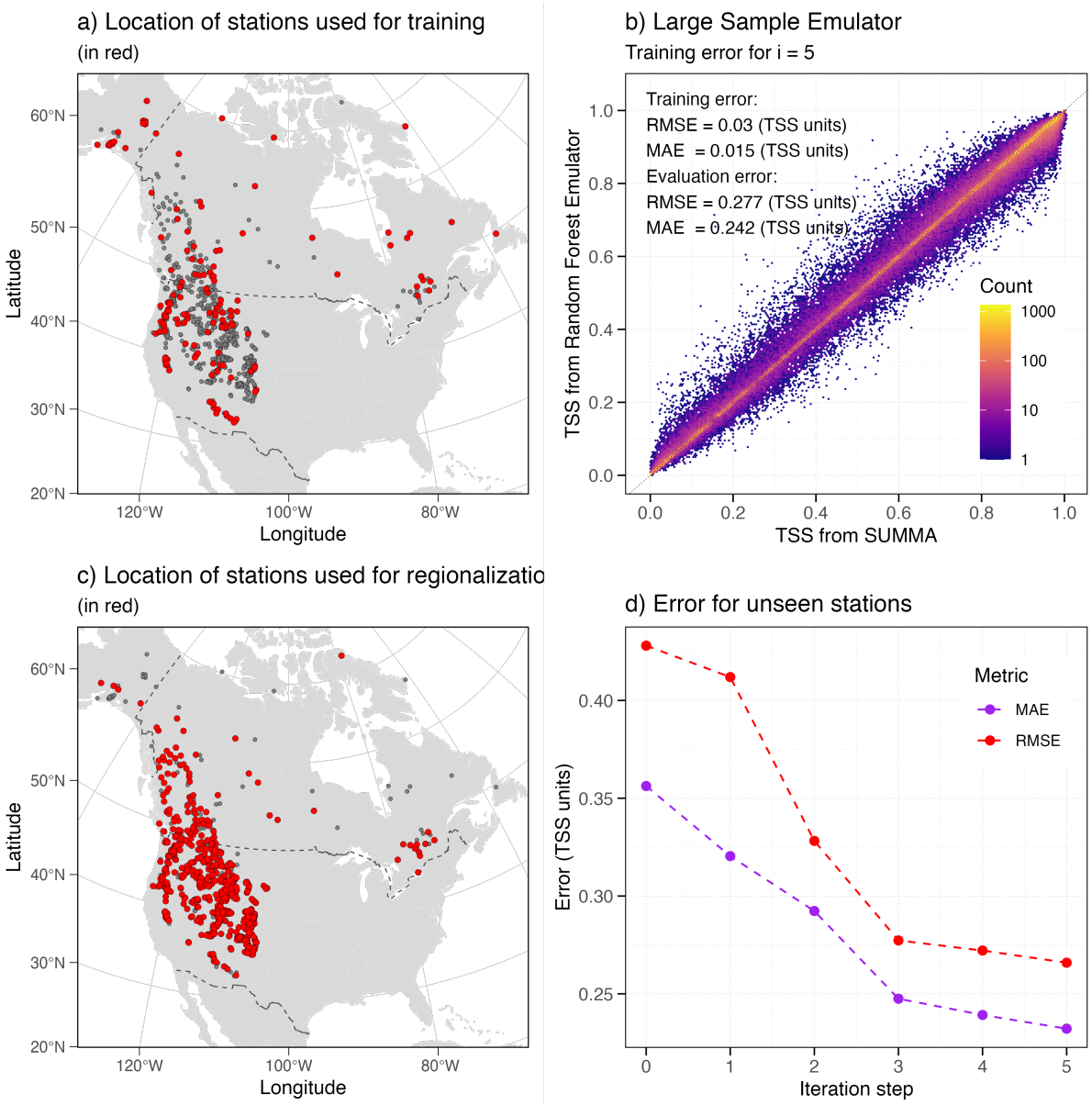


Figure C1. Emulator errors for training and unseen stations. a) location of the basins (in red) used to train the emulator (n = 162). b) Training error for emulator at iteration step $i = 5$, and evaluation errors for the new parameter sets obtained through the Genetic Algorithm optimization search. c) stations used for parameter regionalization (i.e., unseen stations, n = 608). d) Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) between SUMMA TSS and emulated TSS for the unseen stations and from the emulator trained for each step of the iterative process to refine the emulator.

Appendix D: Spatially Distributed Snow Measurement and Precipitation Phase Results

Spatially distributed results are provided in this appendix to illustrate how spatial patterns and biases are present. This is presented first for the comparison of model results to MODIS derived snow metrics in Figure D1 and D2 followed by assessment of precipitation phase uncertainty in Figure D3.

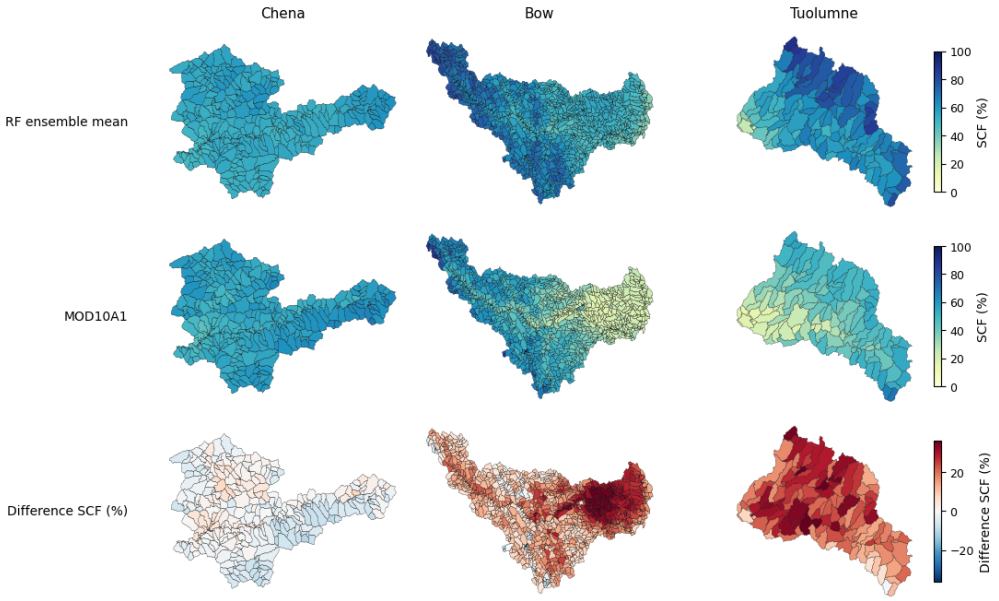


Figure D1. Snow Cover Fraction (SCF) averaged over all water years (2010-2018). This plot show the RF ensemble mean result, the MODIS measurement and difference between. Snow Cover Fraction (SCF) is overestimated in lower elevation sub-basins in both the Bow and the Tuolumne. This would contribute to a overestimation in catchment mean assessment. Chena results are more consistent spatially and with closer agreement between model and measurement

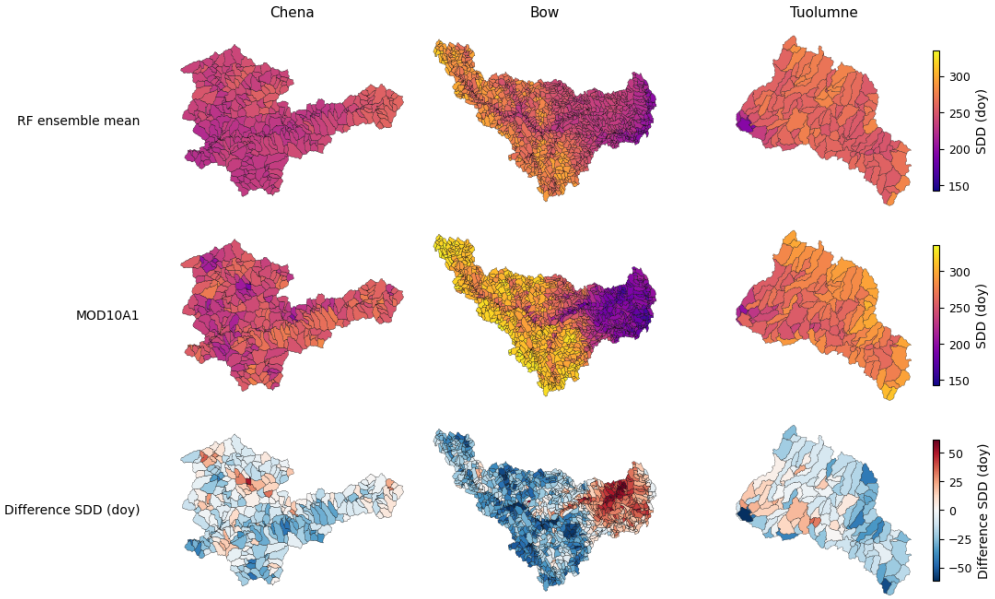


Figure D2. Snow Disappearance Date (SDD) averaged over all water years (2010-2018). This plot show the RF ensemble mean result, the MODIS measurement and difference between. Comparable to the SCF result, the SDD is overestimated in the lower elevations for Bow and Tuolumne while the upper elevation sub-basin SDD tends to be more underestimated by the model.

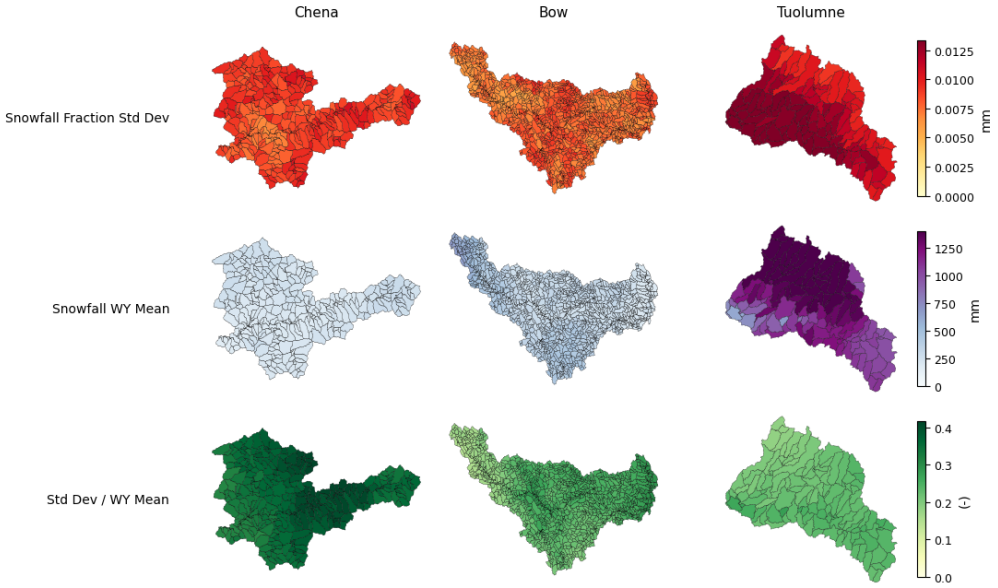


Figure D3. Snowfall fraction is calculated across ensemble members and the standard deviation used to represent precipitation phase uncertainty. The uncertainty in precipitation phase across the simulation period does not indicate a distinct spatial pattern, with the exception of high uncertainty at Tuolumne lower elevations. Normalizing the standard deviation by the total water years (WY) mean allows comparison across basins.

Appendix E: Complete SWE Timeseries Results

The complete SWE timeseries allowing comparison between modeled and measured results is shown in Figure E1, E2 and E3.

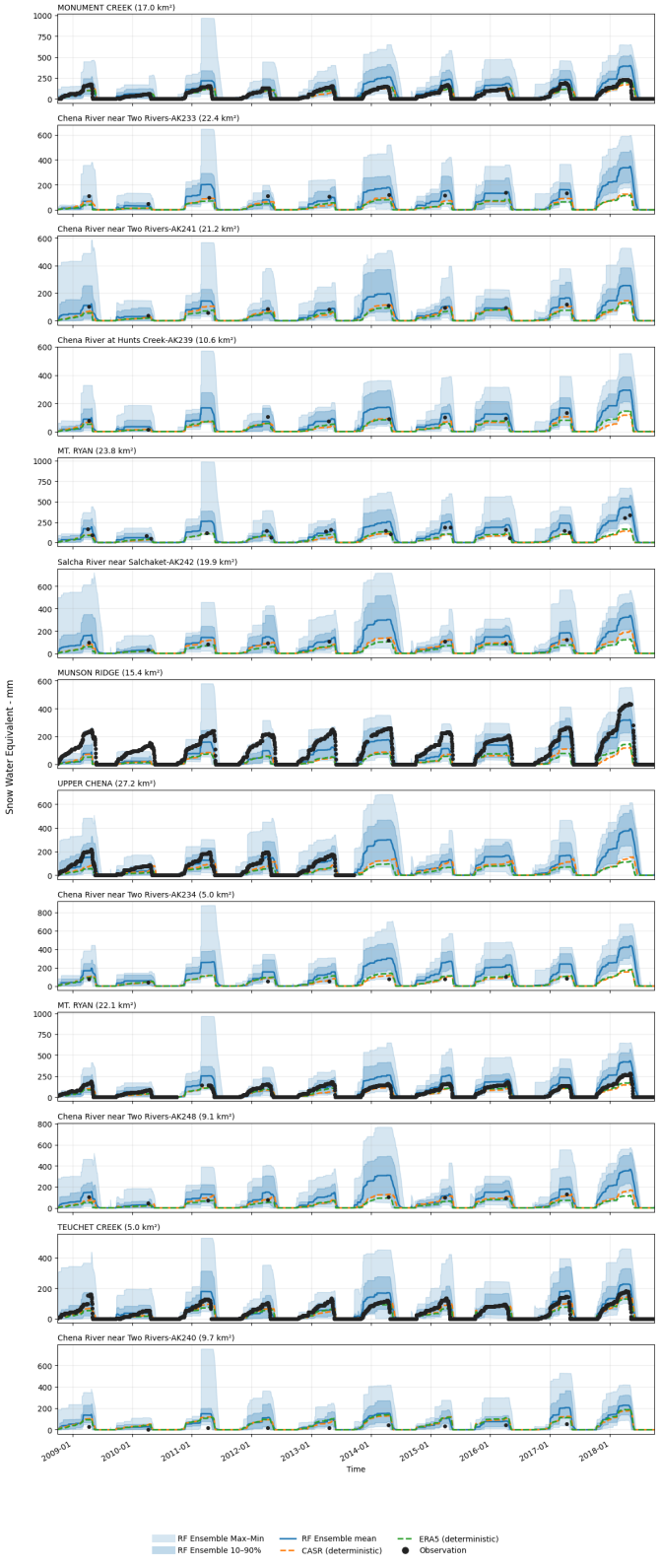


Figure E1. Chena SUMMA Snow Simulation Timeseries compared to NorSWE observations

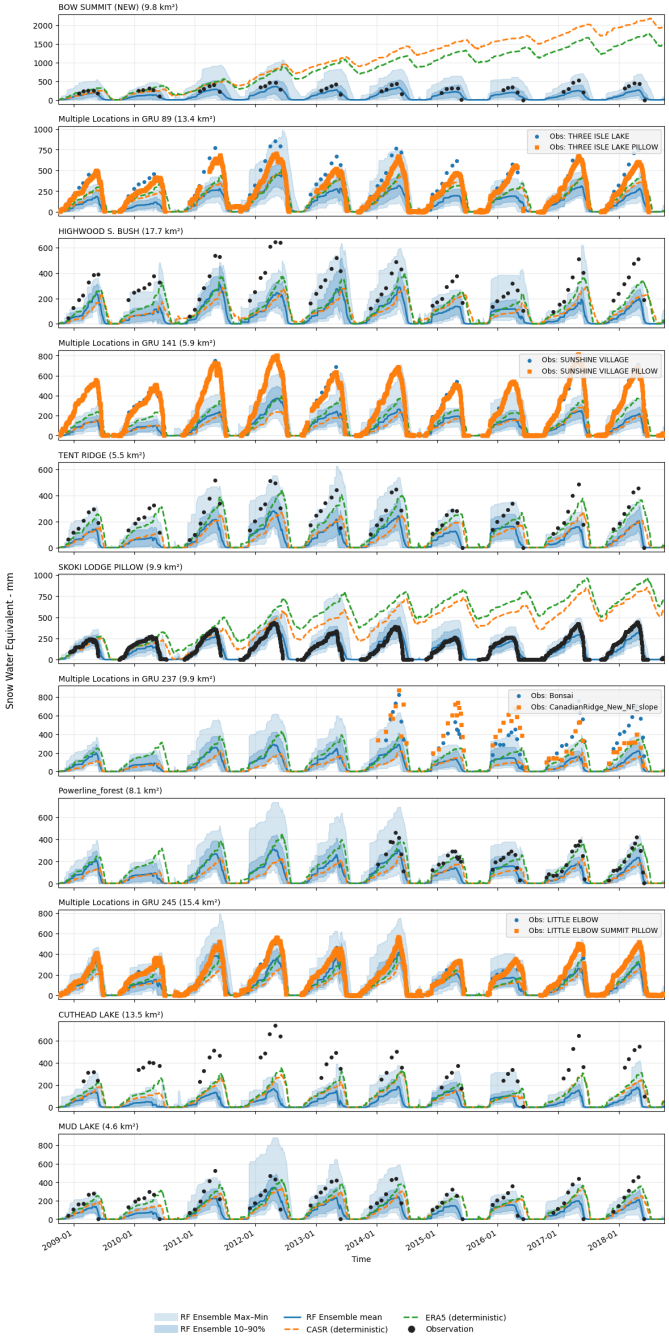


Figure E2. Bow SUMMA Snow Simulation Timeseries compared to NorSWE observations

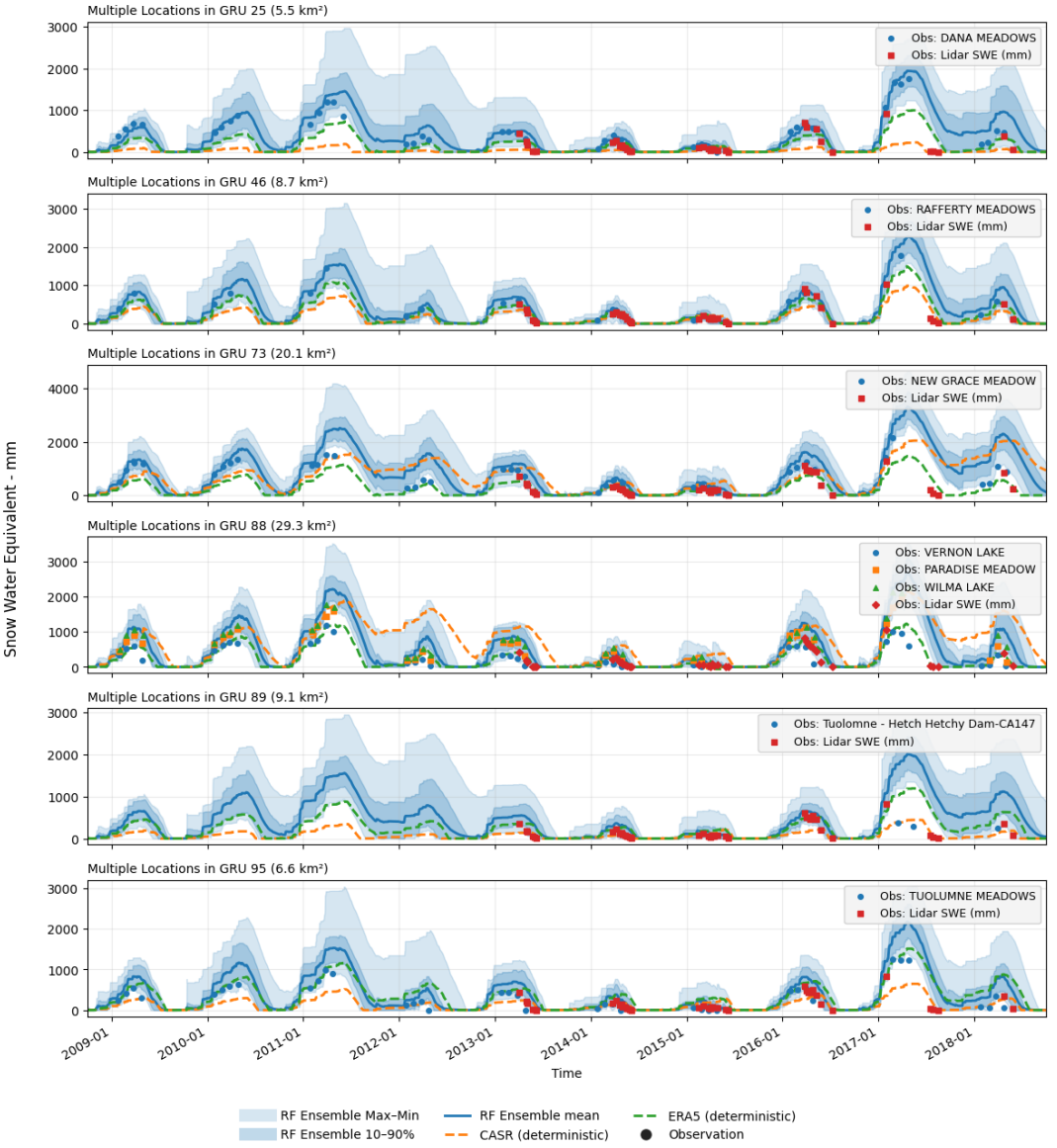


Figure E3. Tuolumne SUMMA Snow Simulation Timeseries compared to NorSWE observations. ASO lidar estimates of SWE are also shown as sub-basin averages from 2014-2018

Author contributions. DC led the research, manuscript writing and performed the majority of technical tasks (Conceptualization; Methodology; Software; Data curation; Formal analysis; Investigation; Validation; Visualization; Writing, original draft). GT provided GPEP software guidance and helped shape the study approach (Software; Methodology; Conceptualization; Writing, review and editing). NV designed and executed SUMMA snow calibration, contributed to evaluation and figures, and reviewed the manuscript (Methodology; Software; For-

mal analysis; Investigation; Validation; Visualization; Writing, review and editing). AW contributed initial SUMMA model formulations, implementation code, and GPEP-prototype datasets (GMET) for the Tuolumne Basin, and to study design and manuscript revisions (Conceptualization; Software; Methodology; Writing, review and editing). MC supervised the research and secured funding (Supervision; Funding acquisition; Project administration; Conceptualization; Writing, review and editing). All authors reviewed and approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Wouter Knoben and Gaby Gründemann for insights on scientific workflows and for constructive manuscript edits that improved this work. This research was supported by the Cooperative Institute for Research to Operations in Hydrology (CIROH) with funding under award NA22NWS4320003 from the NOAA Cooperative Institute Program. Co-author AW was also supported by funding from the US Army Corps of Engineers and the US Bureau of Reclamation Science and Technology program. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the opinions of the supporting agencies.

Portions of this work were supported by the use of AI tools. In particular, ChatGPT 5.1 (OpenAI) was used to assist in improving the clarity and consistency of the manuscript text, Latex formatting, and to support the design and refinement of figures. The scientific content, analyses, and interpretations are the responsibility of the authors, who reviewed and approved all AI-assisted contributions.

References

Anderson, G. S.: Hydrologic Reconnaissance of the Tanana Basin, Central Alaska, Hydrologic Atlas, https://doi.org/10.3133/ha319, 1970.

Arnal, L., Clark, M. P., Pietroniro, A., Vionnet, V., Casson, D. R., Whitfield, P. H., Fortin, V., Wood, A. W., Knoben, W. J. M., Newton, B. W., and Walford, C.: FROSTBYTE: A Reproducible Data-Driven Workflow for Probabilistic Seasonal Streamflow Forecasting in Snow-Fed River Basins across North America, Hydrology and Earth System Sciences, 28, 4127–4155, https://doi.org/10.5194/hess-28-4127-2024, 2024.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J. D., and Xuan Thinh, N.: RF-MEP: A Novel Random Forest Method for Merging Gridded Precipitation Products and Ground-Based Measurements, Remote Sensing of Environment, 239, 111 606, https://doi.org/10.1016/j.rse.2019.111606, 2020.

Barlage, M., Chen, F., Tewari, M., Ikeda, K., Gochis, D., Dudhia, J., Rasmussen, R., Livneh, B., Ek, M., and Mitchell, K.: Noah Land Surface Model Modifications to Improve Snowpack Prediction in the Colorado Rocky Mountains, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD013470, 2010.

Bash, E. A. and Marshall, S. J.: Estimation of Glacial Melt Contributions to the Bow River, Alberta, Canada, Using a Radiation-Temperature Melt Model, Annals of Glaciology, 55, 138–152, https://doi.org/10.3189/2014AoG66A226, 2014.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-Scale Evaluation of 23 Precipitation Datasets Using Gaugeobservations and Hydrological Modeling, https://doi.org/10.5194/hess-2017-508, 2017.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Behrangi, A., Gardner, A., Reager, J. T., Fisher, J. B., Yang, D., Huffman, G. J., and Adler, R. F.: Using GRACE to Estitmate Snowfall Accumulation and Assess Gauge Undercatch Corrections in High Latitudes, https://doi.org/10.1175/JCLI-D-18-0163.1, 2018.

Bennett, A. R., Hamman, J. J., and Nijssen, B.: MetSim: A Python Package for Estimation and Disaggregation of Meteorological Data, Journal of Open Source Software, 5, 2042, https://doi.org/10.21105/joss.02042, 2020.

Bennett, K. E., Cherry, J. E., Balk, B., and Lindsey, S.: Using MODIS Estimates of Fractional Snow Cover Area to Improve Streamflow Forecasts in Interior Alaska, Hydrology and Earth System Sciences, 23, 2439–2459, https://doi.org/10.5194/hess-23-2439-2019, 2019.

Berg, J., Reynolds, D., Quéno, L., Jonas, T., Lehning, M., and Mott, R.: A Seasonal Snowpack Model Forced with Dynamically Downscaled Forcing Data Resolves Hydrologically Relevant Accumulation Patterns, Frontiers in Earth Science, 12, https://doi.org/10.3389/feart.2024.1393260, 2024.

Bertoncini, A. and Pomeroy, J. W.: Quantifying Spatiotemporal and Elevational Precipitation Gauge Network Uncertainty in the Canadian Rockies, Hydrology and Earth System Sciences, 29, 983–1000, https://doi.org/10.5194/hess-29-983-2025, 2025.

Blöschl, G. and Sivapalan, M.: Scale Issues in Hydrological Modelling: A Review, Hydrological Processes, 9, 251–290, https://doi.org/10.1002/hyp.3360090305, 1995.

Bohn, T. J., Livneh, B., Oyler, J. W., Running, S. W., Nijssen, B., and Lettenmaier, D. P.: Global Evaluation of MTCLIM and Related Algorithms for Forcing of Ecological and Hydrological Models, Agricultural and Forest Meteorology, 176, 38–49, https://doi.org/10.1016/j.agrformet.2013.03.003, 2013.

Böhner, J. and Antonić, O.: Chapter 8 Land-Surface Parameters Specific to Topo-Climatology, in: Developments in Soil Science, edited by Hengl, T. and Reuter, H. I., vol. 33 of *Geomorphometry*, pp. 195–226, Elsevier, https://doi.org/10.1016/S0166-2481(08)00008-1, 2009.

BRCB: Bow River Basin State of the Watershed Summary, 2010, Bow River Council, Bow River Watershed (Alta.), ISBN 20118003283, 2010.

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

Brown, R. D., Fang, B., and Mudryk, L.: Update of Canadian Historical Snow Survey Data and Analysis of Snow Water Equivalent Trends, 1967–2016, Atmosphere-Ocean, 57, 149–156, https://doi.org/10.1080/07055900.2019.1598843, 2019.

Bunn, P. T., Wood, A. W., Newman, A. J., Chang, H.-I., Castro, C. L., Clark, M. P., and Arnold, J. R.: Improving Station-Based Ensemble Surface Meteorological Analyses Using Numerical Weather Prediction: A Case Study of the Oroville Dam Crisis Precipitation Event, Journal of Hydrometeorology, https://doi.org/10.1175/JHM-D-21-0193.1, 2022.

Carlson, K. A., Levin, H. K., Morris, A. L., Candela, S. G., Rivera, A. M. M., Huening, V. G., and Fredericks, J. G.: TDX-hydro: Global High-Resolution Hydrography Derived from TanDEM-x, https://doi.org/10.22541/essoar.171629686.65893579/v1, 2024.

Carlstein, E.: The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence, The Annals of Statistics, 14, 1171–1179, 1986.

Casson, D. R.: Ensemble Meteorological Forcing for Chena, Bow and Tuolumne, https://doi.org/10.20383/103.01532, 2025a.

Casson, D. R.: Gpep_snakemake: Workflow for Generating Probabilistic Ensemble Meteorological Forcings, https://doi.org/10.5281/zenodo.17727619, 2025b.

Casson, D. R.: Gpep_to_summa_snakemake: Workflow for Coupling GPEP Ensemble Forcings to the Physically Based Hydrological Modelling, https://doi.org/10.5281/zenodo.17727623, 2025c.

Casson, D. R.: SUMMA Model and GPEP Input Files, https://doi.org/10.5281/zenodo.17740654, 2025d.

Casson, D. R., Werner, M., Weerts, A., and Solomatine, D.: Global Re-Analysis Datasets to Improve Hydrological Assessment and Snow Water Equivalent Estimation in a Sub-Arctic Watershed, Hydrology and Earth System Sciences, 22, 4685–4697, https://doi.org/10.5194/hess-22-4685-2018, 2018.

Chaney, N. W., Van Huijgevoort, M. H. J., Shevliakova, E., Malyshev, S., Milly, P. C. D., Gauthier, P. P. G., and Sulman, B. N.: Harnessing Big Data to Rethink Land Heterogeneity in Earth System Models, Hydrology and Earth System Sciences, 22, 3311–3330, https://doi.org/10.5194/hess-22-3311-2018, 2018.

Clark, M. P. and Slater, A. G.: Probabilistic Quantitative Precipitation Estimation in Complex Terrain, Journal of Hydrometeorology, 7, 3–22, https://doi.org/10.1175/JHM474.1, 2006.

Clark, M. P., Slater, A. G., Barrett, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., and Leavesley, G. H.: Assimilation of Snow Covered Area Information into Hydrologic and Land-Surface Models, Advances in Water Resources, 29, 1209–1221, https://doi.org/10.1016/j.advwatres.2005.10.001, 2006.

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Hreinsson, E. Ö., and Woods, R. A.: Representing Spatial Variability of Snow Water Equivalent in Hydrologic and Land-Surface Models: A Review, Water Resources Research, 47, https://doi.org/10.1029/2011WR010745, 2011.

Clark, M. P., Nijssen, B., Lundquist, D., Kavetski, D., Rupp, E., Woods, A., Freer, E., Gutmann, D., Wood, W., Brekke, D., Arnold, R., Gochis, J., Rasmussen, M., Tarboton, G., Mahat, V., Flerchinger, N., and Marks, G.: The Structure for Unifying Multiple Modeling Alternatives (SUMMA), Version 1.0: Technical Description, https://doi.org/10.5065/D6WQ01TD, 2015a.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A Unified Approach for Process-Based Hydrologic Modeling: 1. Modeling Concept, Water Resources Research, 51, 2498–2514, https://doi.org/10.1002/2015WR017198, 2015b.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., Rasmussen, R. M., Tarboton, D. G., Mahat, V., Flerchinger, G. N., and Marks, D. G.: A Unified Approach for Process-Based Hydrologic Modeling: 2. Model Implementation and Case Studies, Water Resources Research, 51, 2515–2542, https://doi.org/10.1002/2015WR017200, 2015c.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029 001, https://doi.org/10.1029/2020WR029001, 2021a.

Clark, M. P., Zolfaghari, R., Green, K. R., Trim, S., Knoben, W. J. M., Bennett, A., Nijssen, B., Ireson, A., and Spiteri, R. J.: The Numerical Implementation of Land Models: Problem Formulation and Laugh Tests, Journal of Hydrometeorology, 22, 1627–1648, https://doi.org/10.1175/JHM-D-20-0175.1, 2021b.

Cloke, H. L. and Pappenberger, F.: Ensemble Flood Forecasting: A Review, Journal of Hydrology, 375, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005, 2009.

Colli, M., Pollock, M., Stagnaro, M., Lanza, L. G., Dutton, M., and O'Connell, E.: A Computational Fluid-Dynamics Assessment of the Improved Performance of Aerodynamic Rain Gauges, Water Resources Research, 54, 779–796, https://doi.org/10.1002/2017WR020549, 2018.

Colli, M., Stagnaro, M., Lanza, L. G., Rasmussen, R., and Thériault, J. M.: Adjustments for Wind-Induced Undercatch in Snowfall Measurements Based on Precipitation Intensity, https://doi.org/10.1175/JHM-D-19-0222.1, 2020.

Collins, M.: Ensembles and Probabilities: A New Era in the Prediction of Climate Change, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 1957–1970, https://doi.org/10.1098/rsta.2007.2068, 2007.

Contractor, S., Donat, M. G., Alexander, L. V., Ziese, M., Meyer-Christoffer, A., Schneider, U., Rustemeier, E., Becker, A., Durre, I., and Vose, R. S.: Rainfall Estimates on a Gridded Network (REGEN) – a Global Land-Based Gridded Dataset of Daily Precipitation from 1950 to 2016, Hydrology and Earth System Sciences, 24, 919–943, https://doi.org/10.5194/hess-24-919-2020, 2020.

Cristea, N. C., Lundquist, J. D., Loheide II, S. P., Lowry, C. S., and Moore, C. E.: Modelling How Vegetation Cover Affects Climate Change Impacts on Streamflow Timing and Magnitude in the Snowmelt-Dominated Upper Tuolumne Basin, Sierra Nevada, Hydrological Processes, 28, 3896–3918, https://doi.org/10.1002/hyp.9909, 2014.

Crumley, R. L., Palomaki, R. T., Nolin, A. W., Sproles, E. A., and Mar, E. J.: SnowCloudMetrics: Snow Information for Everyone, Remote Sensing, 12, 3341, https://doi.org/10.3390/rs12203341, 2020.

Currier, W. R. and Lundquist, J. D.: Snow Depth Variability at the Forest Edge in Multiple Climates in the Western United States, Water Resources Research, 54, 8756–8773, https://doi.org/10.1029/2018WR022553, 2018.

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J.: Random Forests for Classification in Ecology, Ecology, 88, 2783–2792, https://doi.org/10.1890/07-0539.1, 2007.

Daly, C., Neilson, R. P., and Phillips, D. L.: A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain, 1994.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically Sensitive Mapping of Climatological Temperature and Precipitation across the Conterminous United States, International Journal of Climatology, 28, 2031–2064, https://doi.org/10.1002/joc.1688, 2008.

Daly, C., Doggett, M. K., Smith, J. I., Olson, K. V., Halbleib, M. D., Dimcovic, Z., Keon, D., Loiselle, R. A., Steinberg, B., Ryan, A. D., Pancake, C. M., and Kaspar, E. M.: Challenges in Observation-Based Mapping of Daily Precipitation across the Conterminous United States, https://doi.org/10.1175/JTECH-D-21-0054.1, 2021.

Daly, S. F. and Vuyovich, C. M.: The Chena River Watershed Hydrology Model, Technical Report ERDC/CRREL TR-12-1, U.S.Army Engineer Research and Development Center, Cold Regions Research and Engineering Laboratory (ERDC/CRREL), 2012.

Deems, J. S., Painter, T. H., and Finnegan, D. C.: Lidar Measurement of Snow Depth: A Review, Journal of Glaciology, 59, 467–479, 2013.

Dozier, J., Bair, E. H., and Davis, R. E.: Estimating the Spatial Distribution of Snow Water Equivalent in the World's Mountains, WIREs Water, 3, 461–474, https://doi.org/10.1002/wat2.1140, 2016.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S.: Comprehensive Automated Quality Assurance of Daily Surface Observations, https://doi.org/10.1175/2010JAMC2375.1, 2010.

Ensor, L. A. and Robeson, S. M.: Statistical Characteristics of Daily Precipitation: Comparisons of Gridded and Point Datasets, https://doi.org/10.1175/2008JAMC1757.1, 2008.

Essery, R., Morin, S., Lejeune, Y., and B Ménard, C.: A Comparison of 1701 Snow Models Using Observations from an Alpine Site, Advances in Water Resources, 55, 131–148, https://doi.org/10.1016/j.advwatres.2012.07.013, 2013.

Fang, X., Pomeroy, J. W., DeBeer, C. M., Harder, P., and Siemens, E.: Hydrometeorological Data from Marmot Creek Research Basin, Canadian Rockies, Earth System Science Data, 11, 455–471, https://doi.org/10.5194/essd-11-455-2019, 2019.

Farahani, M. A., Wood, A. W., Tang, G., and Mizukami, N.: Calibrating a Large-Domain Land/Hydrology Process Model in the Age of AI: The SUMMA CAMELS Emulator Experiments, Hydrology and Earth System Sciences, 29, 4515–4537, https://doi.org/10.5194/hess-29-4515-2025, 2025.

Fleming, S. W., Zukiewicz, L., Strobel, M. L., Hofman, H., and Goodbody, A. G.: SNOTEL, the Soil Climate Analysis Network, and Water Supply Forecasting at the Natural Resources Conservation Service: Past, Present, and Future, JAWRA Journal of the American Water Resources Association, 59, 585–599, https://doi.org/10.1111/1752-1688.13104, 2023.

Fox, E. W., Hoef, J. M. V., and Olsen, A. R.: Comparing Spatial Regression to Random Forests for Large Environmental Data Sets, PLOS ONE, 15, e0229 509, https://doi.org/10.1371/journal.pone.0229509, 2020.

Frei, C. and Isotta, F. A.: Ensemble Spatial Precipitation Analysis From Rain Gauge Data: Methodology and Application in the European Alps, Journal of Geophysical Research: Atmospheres, p. 2018JD030004, https://doi.org/10.1029/2018JD030004, 2019.

Frey, P. J., Mueller, E. W., and Berry, E. C.: The Chena River: The Study of a Subarctic Stream, Federal Water Quality Administration, Department of the Interior, Alaska Water Laboratory, 1970.

Friedl, M. and Sulla-Menashe, D.: MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006, https://doi.org/10.5067/MODIS/MCD12Q1.006, 2019.

Froidurot, S., Zin, I., Hingray, B., and Gautheron, A.: Sensitivity of Precipitation Phase over the Swiss Alps to Different Meteorological Variables, https://doi.org/10.1175/JHM-D-13-073.1, 2014.

Gascoin, S., Luojus, K., Nagler, T., Lievens, H., Masiokas, M., Jonas, T., Zheng, Z., and De Rosnay, P.: Remote Sensing of Mountain Snow from Space: Status and Recommendations, Frontiers in Earth Science, 12, https://doi.org/10.3389/feart.2024.1381323, 2024.

Gasset, N., Fortin, V., Dimitrijevic, M., Carrera, M., Bilodeau, B., Muncaster, R., Gaborit, É., Roy, G., Pentcheva, N., Bulat, M., Wang, X., Pavlovic, R., Lespinas, F., Khedhaouiria, D., and Mai, J.: A 10km North American Precipitation and Land-Surface Reanalysis Based on the GEM Atmospheric Model, Hydrology and Earth System Sciences, 25, 4917–4945, https://doi.org/10.5194/hess-25-4917-2021, 2021.

Gervais, M., Tremblay, L. B., Gyakum, J. R., and Atallah, E.: Representing Extremes in a Daily Gridded Precipitation Analysis over the United States: Impacts of Station Density, Resolution, and Gridding Methods, https://doi.org/10.1175/JCLI-D-13-00319.1, 2014.

Gharari, S., Keshavarz, K., Knoben, W. J. M., Tang, G., and Clark, M. P.: EASYMORE: A Python Package to Streamline the Remapping of Variables for Earth System Models, SoftwareX, 24, https://doi.org/10.1016/j.softx.2023.101547, 2023.

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J.-H., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., and Yu, X.: Toward the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance, Earth and Space Science, 3, 388–415, https://doi.org/10.1002/2015EA000136, 2016.

Girotto, M., Musselman, K. N., and Essery, R. L. H.: Data Assimilation Improves Estimates of Climate-Sensitive Seasonal Snow, Current Climate Change Reports, 6, 81–94, https://doi.org/10.1007/s40641-020-00159-7, 2020.

Goodison, B. E., Louie, P. Y. T., and Yang, D.: WMO Solid precipitation measurement intercomparison - Final report, https://library.wmo.int/es/records/item/28336-wmo-solid-precipitation-measurement-intercomparison, 1998.

Günther, D., Marke, T., Essery, R., and Strasser, U.: Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing Data Error on Point-Scale Energy Balance Snow Model Performance, Water Resources Research, 55, 2779–2800, https://doi.org/10.1029/2018WR023403, 2019.

Hall, D. K. and Riggs, G. A.: MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, Version 61, https://doi.org/10.5067/MODIS/MOD10A1.061, 2021.

Hallouin, T., Bourgin, F., Perrin, C., Ramos, M.-H., and Andréassian, V.: EvalHyd v0.1.2: A Polyglot Tool for the Evaluation of Deterministic and Probabilistic Streamflow Predictions, Geoscientific Model Development, 17, 4561–4578, https://doi.org/10.5194/gmd-17-4561-2024, 2024.

Hedrick, A. R., Marks, D., Marshall, H.-P., McNamara, J., Havens, S., Trujillo, E., Sandusky, M., Robertson, M., Johnson, M., Bormann, K. J., and Painter, T. H.: From Drought to Flood: A Water Balance Analysis of the Tuolumne River Basin during Extreme Conditions (2015–2017), Hydrological Processes, 34, 2560–2574, https://doi.org/10.1002/hyp.13749, 2020.

Hedstrom, N. R. and Pomeroy, J. W.: Measurements and Modelling of Snow Interception in the Boreal Forest, Hydrological Processes, 12, 1611–1625, https://doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1611::AID-HYP684>3.0.CO;2-4, 1998.

Helbig, N., Mott, R., van Herwijnen, A., Winstral, A., and Jonas, T.: Parameterizing Surface Wind Speed over Complex Topography, Journal of Geophysical Research: Atmospheres, 122, 651–667, https://doi.org/10.1002/2016JD025593, 2017.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global Gridded Soil Information Based on Machine Learning, PLOS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Henn, B., Clark, M. P., Kavetski, D., McGurk, B., Painter, T. H., and Lundquist, J. D.: Combining Snow, Streamflow, and Precipitation Gauge Observations to Infer Basin-Mean Precipitation, Water Resources Research, 52, 8700–8723, https://doi.org/10.1002/2015WR018564, 2016.

Henn, B., Clark, M. P., Kavetski, D., Newman, A. J., Hughes, M., McGurk, B., and Lundquist, J. D.: Spatiotemporal Patterns of Precipitation Inferred from Streamflow Observations across the Sierra Nevada Mountain Range, Journal of Hydrology, 556, 993–1012, https://doi.org/10.1016/j.jhydrol.2016.08.009, 2018a.

Henn, B., Newman, A. J., Livneh, B., Daly, C., and Lundquist, J. D.: An Assessment of Differences in Gridded Precipitation Datasets in Complex Terrain, Journal of Hydrology, 556, 1205–1219, https://doi.org/10.1016/j.jhydrol.2017.03.008, 2018b.

Henn, B., Painter, T. H., Bormann, K. J., McGurk, B., Flint, A. L., Flint, L. E., White, V., and Lundquist, J. D.: High-Elevation Evapotranspiration Estimates During Drought: Using Streamflow and NASA Airborne Snow Observatory SWE Observations to Close the Upper Tuolumne River Basin Water Balance, Water Resources Research, 54, 746–766, https://doi.org/10.1002/2017WR020473, 2018c.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hong, Y., Xuan Do, H., Kessler, J., Fry, L., Read, L., Rafieei Nasab, A., Gronewold, A. D., Mason, L., and Anderson, E. J.: Evaluation of Gridded Precipitation Datasets over International Basins and Large Lakes, Journal of Hydrology, 607, 127 507, https://doi.org/10.1016/j.jhydrol.2022.127507, 2022.

Hotovy, O., Nedelcev, O., Seibert, J., and Jenicek, M.: Rain-on-Snow Events in Mountainous Catchments under Climate Change, Hydrology and Earth System Sciences, 29, 4199–4217, https://doi.org/10.5194/hess-29-4199-2025, 2025.

Hrachowitz, M. and Clark, M. P.: HESS Opinions: The Complementary Merits of Competing Modelling Philosophies in Hydrology, Hydrology and Earth System Sciences, 21, 3953–3973, https://doi.org/10.5194/hess-21-3953-2017, 2017.

Huidobro, G., Chiu, C.-M., Byun, K., and Hamlet, A. F.: Meteorological Driving Datasets for the U.S. Midwest and Great Lakes Region Incorporating Precipitation Gauge Undercatch Corrections, https://doi.org/10.1175/JHM-D-22-0148.1, 2023.

Hungerford, R. D., Nemani, R. R., Running, S. W., and Coughlan, J. C.: MTCLIM: A Mountain Microclimate Simulation Model, Tech. Rep. INT-RP-414, U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT, https://doi.org/10.2737/INT-RP-414, 1989.

Huning, L. S. and AghaKouchak, A.: Global Snow Drought Hot Spots and Characteristics, Proceedings of the National Academy of Sciences, 117, 19 753–19 759, https://doi.org/10.1073/pnas.1915921117, 2020.

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., Hyde, S., Brumby, S., Davies, B. J., Elmore, A. C., Emmer, A., Feng, M., Fernández, A., Haritashya, U., Kargel, J. S., Koppes, M., Kraaijenbrink, P. D. A., Kulkarni, A. V., Mayewski, P. A., Nepal, S., Pacheco, P., Painter, T. H., Pellicciotti, F., Rajaram, H., Rupper, S., Sinisalo, A., Shrestha, A. B., Viviroli, D., Wada, Y., Xiao, C., Yao, T., and Baillie, J. E. M.: Importance and Vulnerability of the World's Water Towers, Nature, 577, 364–369, https://doi.org/10.1038/s41586-019-1822-y, 2020.

Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A. M., Helber, R., and Arnone, R. A.: Summary Diagrams for Coupled Hydrodynamic-Ecosystem Model Skill Assessment, Journal of Marine Systems, 76, 64–82, https://doi.org/10.1016/j.jmarsys.2008.05.014, 2009.

Jordan, R.: A One-Dimensional Temperature Model for a Snow Cover, Special report (Cold Regions Research and Engineering Laboratory (U.S.)), 1991.

Kim, R. S., Kumar, S., Vuyovich, C., Houser, P., Lundquist, J., Mudryk, L., Durand, M., Barros, A., Kim, E. J., Forman, B. A., Gutmann, E. D., Wrzesien, M. L., Garnaud, C., Sandells, M., Marshall, H.-P., Cristea, N., Pflug, J. M., Johnston, J., Cao, Y., Mocko, D., and Wang, S.:

Snow Ensemble Uncertainty Project (SEUP): Quantification of Snow Water Equivalent Uncertainty across North America via Ensemble Land Surface Modeling, The Cryosphere, 15, 771–791, https://doi.org/10.5194/tc-15-771-2021, 2021.

Kimball, J. S., Running, S. W., and Nemani, R.: An Improved Method for Estimating Surface Humidity from Daily Minimum Temperature, Agricultural and Forest Meteorology, 85, 87–98, https://doi.org/10.1016/S0168-1923(96)02366-0, 1997.

Knoben, W.: Global USDA-NRCS Soil Texture Class Map, https://doi.org/10.4211/hs.1361509511e44adfba814f6950c6e742, 2021.

Knoben, W. J. M., Clark, M. P., Bales, J., Bennett, A., Gharari, S., Marsh, C. B., Nijssen, B., Pietroniro, A., Spiteri, R. J., Tang, G., Tarboton, D. G., and Wood, A. W.: Community Workflows to Advance Reproducibility in Hydrologic Modeling: Separating Model-Agnostic and Model-Specific Configuration Steps in Applications of Large-Domain Hydrologic Models, Water Resources Research, 58, e2021WR031 753, https://doi.org/10.1029/2021WR031753, 2022.

Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K., Smith, C. D., Yang, D., Roulet, Y.-A., Buisan, S., Laine, T., Lee, G., Aceituno, J. L. C., Alastrué, J., Isaksen, K., Meyers, T., Brækkan, R., Landolt, S., Jachcik, A., and Poikonen, A.: Analysis of Single-Alter-shielded and Unshielded Measurements of Mixed and Solid Precipitation from WMO-SPICE, Hydrology and Earth System Sciences, 21, 3525–3542, https://doi.org/10.5194/hess-21-3525-2017, 2017a.

Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E., Meyers, T., Landolt, S., Jachcik, A., Isaksen, K., Brækkan, R., and Leeper, R.: The Quantification and Correction of Wind-Induced Precipitation Measurement Errors, Hydrology and Earth System Sciences, 21, 1973–1989, https://doi.org/10.5194/hess-21-1973-2017, 2017b.

Köster, J. and Rahmann, S.: Snakemake—a Scalable Bioinformatics Workflow Engine, Bioinformatics, 28, 2520–2522, https://doi.org/10.1093/bioinformatics/bts480, 2012.

Largeron, C., Dumont, M., Morin, S., Boone, A., Lafaysse, M., Metref, S., Cosme, E., Jonas, T., Winstral, A., and Margulis, S. A.: Toward Snow Cover Estimation in Mountainous Areas Using Modern Data Assimilation Methods: A Review, Frontiers in Earth Science, 8, https://doi.org/10.3389/feart.2020.00325, 2020.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G. B., and Slater, A. G.: Parameterization Improvements and Functional and Structural Advances in Version 4 of the Community Land Model, Journal of Advances in Modeling Earth Systems, 3, https://doi.org/10.1029/2011MS00045, 2011.

Liao, C. and Zhuang, Q.: Quantifying the Role of Snowmelt in Stream Discharge in an Alaskan Watershed: An Analysis Using a Spatially Distributed Surface Hydrology Model, Journal of Geophysical Research: Earth Surface, 122, 2183–2195, https://doi.org/10.1002/2017JF004214, 2017.

Liu, H., Wood, A. W., Newman, A. J., and Clark, M. P.: Ensemble Dressing of Meteorological Fields: Using Spatial Regression to Estimate Uncertainty in Deterministic Gridded Meteorological Datasets, https://doi.org/10.1175/JHM-D-21-0176.1, 2022.

Lundquist, J., Abel, M. R., Gutmann, E., and Kapnick, S.: Our Skill in Modeling Mountain Rain and Snow Is Bypassing the Skill of Our Observational Networks, Bulletin of the American Meteorological Society, 100, 2473–2490, https://doi.org/10.1175/BAMS-D-19-0001.1, 2019.

Lundquist, J. D., Abel, M. R., Henn, B., Gutmann, E. D., Livneh, B., Dozier, J., and Neiman, P.: High-Elevation Precipitation Patterns: Using Snow Measurements to Assess Daily Gridded Datasets across the Sierra Nevada, California*, Journal of Hydrometeorology, 16, 1773–1792, https://doi.org/10.1175/JHM-D-15-0019.1, 2015a.

Lundquist, J. D., Wayand, N. E., Massmann, A., Clark, M. P., Lott, F., and Cristea, N. C.: Diagnosis of Insidious Data Disasters, Water Resources Research, 51, 3815–3827, https://doi.org/10.1002/2014WR016585, 2015b.

Lundquist, J. D., Roche, J. W., Forrester, H., Moore, C., Keenan, E., Perry, G., Cristea, N., Henn, B., Lapo, K., McGurk, B., Cayan, D. R., and Dettinger, M. D.: Yosemite Hydroclimate Network: Distributed Stream and Atmospheric Data for the Tuolumne River Watershed and Surroundings, Water Resources Research, 52, 7478–7489, https://doi.org/10.1002/2016WR019261, 2016.

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., and Andersen, J.: seNorge2 Daily Precipitation, an Observational Gridded Dataset over Norway from 1957 to the Present Day, Earth System Science Data, 10, 235–249, https://doi.org/10.5194/essd-10-235-2018, 2018.

Magnusson, J., Wever, N., Essery, R., Helbig, N., Winstral, A., and Jonas, T.: Evaluating Snow Models with Varying Process Representations for Hydrological Applications, Water Resources Research, 51, 2707–2723, https://doi.org/10.1002/2014WR016498, 2015.

Margulis, S. A., Fang, Y., Li, D., Lettenmaier, D. P., and Andreadis, K.: The Utility of Infrequent Snow Depth Images for Deriving Continuous Space-Time Estimates of Seasonal Snow Water Equivalent, Geophysical Research Letters, 46, 5331–5340, https://doi.org/10.1029/2019GL082507, 2019.

Marks, D., Kimball, J., Tingey, D., and Link, T.: The Sensitivity of Snowmelt Processes to Climate Conditions and Forest Cover during Rain-on-Snow: A Case Study of the 1996 Pacific Northwest Flood, Hydrological Processes, 12, 1569–1587, https://doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1569::AID-HYP682>3.0.CO;2-L, 1998.

Marsh, C. B., Vionnet, V., and Pomeroy, J. W.: Windmapper: An Efficient Wind Downscaling Method for Hydrological Models, Water Resources Research, 59, e2022WR032 683, https://doi.org/10.1029/2022WR032683, 2023.

Marsh, C. B., Lv, Z., Vionnet, V., Harder, P., Spiteri, R. J., and Pomeroy, J. W.: Snowdrift-Permitting Simulations of Seasonal Snowpack Processes Over Large Mountain Extents, Water Resources Research, 60, e2023WR036 948, https://doi.org/10.1029/2023WR036948, 2024.

Mekis, É. and Vincent, L. A.: An Overview of the Second Generation Adjusted Daily Precipitation Dataset for Trend Analysis in Canada, Atmosphere-Ocean, 49, 163–177, https://doi.org/10.1080/07055900.2011.583910, 2011.

Mekis, E., Donaldson, N., Reid, J., Zucconi, A., Hoover, J., Li, Q., Nitu, R., and Melo, S.: An Overview of Surface-Based Precipitation Observations at Environment and Climate Change Canada, Atmosphere-Ocean, 56, 71–95, https://doi.org/10.1080/07055900.2018.1433627, 2018.

Menard, C. B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A., Brutel-Vuilmet, C., Burke, E., Cuntz, M., Dai, Y., Decharme, B., Dutra, E., Fang, X., Fierz, C., Gusev, Y., Hagemann, S., Haverd, V., Kim, H., Lafaysse, M., Marke, T., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Schädler, G., Semenov, V. A., Smirnova, T., Strasser, U., Swenson, S., Turkov, D., Wever, N., and Yuan, H.: Scientific and Human Errors in a Snow Model Intercomparison, Bulletin of the American Meteorological Society, 102, E61–E79, https://doi.org/10.1175/BAMS-D-19-0329.1, 2021.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, https://doi.org/10.1175/JTECH-D-11-00103.1, 2012.

Miller, Z. S., Peitzsch, E. H., Sproles, E. A., Birkeland, K. W., and Palomaki, R. T.: Assessing the Seasonal Evolution of Snow Depth Spatial Variability and Scaling in Complex Mountain Terrain, The Cryosphere, 16, 4907–4930, https://doi.org/10.5194/tc-16-4907-2022, 2022.

Mortimer, C. and Vionnet, V.: Northern Hemisphere in Situ Snow Water Equivalent Dataset (NorSWE, 1979–2021), Earth System Science Data, 17, 3619–3640, https://doi.org/10.5194/essd-17-3619-2025, 2025.

Mosier, T. M., Hill, D. F., and Sharp, K. V.: How Much Cryosphere Model Complexity Is Just Right? Exploration Using the Conceptual Cryosphere Hydrology Framework, The Cryosphere, 10, 2147–2171, https://doi.org/10.5194/tc-10-2147-2016, 2016.

Mott, R., Vionnet, V., and Grünewald, T.: The Seasonal Snow Cover Dynamics: Review on Wind-Driven Coupling Processes, Frontiers in Earth Science, 6, 197, https://doi.org/10.3389/feart.2018.00197, 2018.

Mott, R., Winstral, A., Cluzet, B., Helbig, N., Magnusson, J., Mazzotti, G., Quéno, L., Schirmer, M., Webster, C., and Jonas, T.: Operational Snow-Hydrological Modeling for Switzerland, Frontiers in Earth Science, 11, https://doi.org/10.3389/feart.2023.1228158, 2023.

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.

Musselman, K. N., Lehner, F., Ikeda, K., Clark, M. P., Prein, A. F., Liu, C., Barlage, M., and Rasmussen, R.: Projected Increases and Shifts in Rain-on-Snow Flood Risk over Western North America, Nature Climate Change, 8, 808–812, https://doi.org/10.1038/s41558-018-0236-4, 2018.

Napoli, A., Crespi, A., Ragone, F., Maugeri, M., and Pasquero, C.: Variability of Orographic Enhancement of Precipitation in the Alpine Region, Scientific Reports, 9, 13 352, https://doi.org/10.1038/s41598-019-49974-5, 2019.

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L., and Arnold, J. R.: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States, Journal of Hydrometeorology, 16, 2481–2500, https://doi.org/10.1175/JHM-D-15-0026.1, 2015.

Newman, A. J., Clark, M. P., Wood, A. W., and Arnold, J. R.: Probabilistic Spatial Meteorological Estimates for Alaska and the Yukon, Journal of Geophysical Research: Atmospheres, 125, e2020JD032 696, https://doi.org/10.1029/2020JD032696, 2020.

Painter, T.: ASO L4 Lidar Snow Water Equivalent 50m UTM Grid, Version 1, https://doi.org/10.5067/M4TUH28NHL4Z, 2018.

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D., Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., Skiles, S. M., Seidel, F. C., and Winstral, A.: The Airborne Snow Observatory: Fusion of Scanning Lidar, Imaging Spectrometer, and Physically-Based Modeling for Mapping Snow Water Equivalent and Snow Albedo, Remote Sensing of Environment, 184, 139–152, https://doi.org/10.1016/j.rse.2016.06.018, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-Learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pflug, J. M. and Lundquist, J. D.: Inferring Distributed Snow Depth by Leveraging Snow Pattern Repeatability: Investigation Using 47 Lidar Observations in the Tuolumne Watershed, Sierra Nevada, California, Water Resources Research, 56, e2020WR027 243, https://doi.org/10.1029/2020WR027243, 2020.

Pierre, A., Jutras, S., Smith, C., Kochendorfer, J., Fortin, V., and Anctil, F.: Evaluation of Catch Efficiency Transfer Functions for Unshielded and Single-Alter-Shielded Solid Precipitation Measurements, Journal of Atmospheric and Oceanic Technology, 36, 865–881, https://doi.org/10.1175/JTECH-D-18-0112.1, 2019.

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., and Carey, S. K.: The Cold Regions Hydrological Model: A Platform for Basing Process Representation and Model Structure on Physical Evidence, Hydrological Processes, 21, 2650–2667, https://doi.org/10.1002/hyp.6787, 2007.

Pomeroy, J. W., Fang, X., and Marks, D. G.: The Cold Rain-on-Snow Event of June 2013 in the Canadian Rockies — Characteristics and Diagnosis, Hydrological Processes, 30, 2899–2914, https://doi.org/10.1002/hyp.10905, 2016.

Raleigh, M. S., Lundquist, J. D., and Clark, M. P.: Exploring the Impact of Forcing Error Characteristics on Physically Based Snow Simulations within a Global Sensitivity Analysis Framework, Hydrology and Earth System Sciences, 19, 3153–3179, 2015.

Rice, R., Bales, R. C., Painter, T. H., and Dozier, J.: Snow Water Equivalent along Elevation Gradients in the Merced and Tuolumne River Basins of the Sierra Nevada, Water Resources Research, 47, https://doi.org/10.1029/2010WR009278, 2011.

Robert, S., Foresti, L., and Kanevski, M.: Spatial Prediction of Monthly Wind Speeds in Complex Terrain with Adaptive General Regression Neural Networks, International Journal of Climatology, 33, 1793–1804, https://doi.org/10.1002/joc.3550, 2013.

Santos, L., Thirel, G., and Perrin, C.: Technical Note: Pitfalls in Using Log-Transformed Flows within the KGE Criterion, Hydrology and Earth System Sciences, 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Schreiner-McGraw, A. P. and Ajami, H.: Combined Impacts of Uncertainty in Precipitation and Air Temperature on Simulated Mountain System Recharge from an Integrated Hydrologic Model, Hydrology and Earth System Sciences, 26, 1145–1164, https://doi.org/10.5194/hess-26-1145-2022, 2022.

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., and Bajat, B.: Random Forest Spatial Interpolation, Remote Sensing, 12, 1687, https://doi.org/10.3390/rs12101687, 2020.

Smith, C. D., Mekis, E., Hartwell, M., and Ross, A.: The Hourly Wind-Bias-Adjusted Precipitation Data Set from the Environment and Climate Change Canada Automated Surface Observation Network (2001–2019), Earth System Science Data, 14, 5253–5265, https://doi.org/10.5194/essd-14-5253-2022, 2022.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., and James, R.: Assessing Data Availability and Research Reproducibility in Hydrology and Water Resources, Scientific Data, 6, 190 030, https://doi.org/10.1038/sdata.2019.30, 2019.

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, https://doi.org/10.1175/2010JHM1202.1, 2010.

Sturm, M., Goldstein, M. A., and Parr, C.: Water and Life from Snow: A Trillion Dollar Science Question, Water Resources Research, 53, 3534–3544, https://doi.org/10.1002/2017WR020840, 2017.

Sun, H. and Margulis, S.: Assessing the Impact of Meteorological Forcing and Its Uncertainty on Snow Modeling and Reanalysis, EGUsphere, pp. 1–34, https://doi.org/10.5194/egusphere-2025-4505, 2025.

Sun, N., Yan, H., Wigmosta, M. S., Leung, L. R., Skaggs, R., and Hou, Z.: Regional Snow Parameters Estimation for Large-Domain Hydrological Applications in the Western United States, Journal of Geophysical Research: Atmospheres, 124, 5296–5313, https://doi.org/10.1029/2018JD030140, 2019.

Tang, G., Clark, M. P., Newman, A. J., Wood, A. W., Papalexiou, S. M., Vionnet, V., and Whitfield, P. H.: SCDNA: A Serially Complete Precipitation and Temperature Dataset for North America from 1979 to 2018, Earth System Science Data, 12, 2381–2409, https://doi.org/10.5194/essd-12-2381-2020, 2020.

Tang, G., Clark, M. P., and Papalexiou, S. M.: SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019, https://doi.org/10.1175/JCLI-D-21-0067.1, 2021.

Tang, G., Wood, A. W., Newman, A. J., Clark, M. P., and Papalexiou, S. M.: GPEP v1.0: The Geospatial Probabilistic Estimation Package to Support Earth Science Applications, Geoscientific Model Development, 17, 1153–1173, https://doi.org/10.5194/gmd-17-1153-2024, 2024.

Tang, G., Clark, M. P., Knoben, W. J. M., Liu, H., Gharari, S., Arnal, L., Wood, A. W., Newman, A. J., Freer, J., and Papalexiou, S. M.: Uncertainty Hotspots in Global Hydrologic Modeling: The Impact of Precipitation and Temperature Forcings, https://doi.org/10.1175/BAMS-D-24-0007.1, 2025a.

Tang, G., Wood, A. W., and Swenson, S.: On Using AI-Based Large-Sample Emulators for Land/Hydrology Model Calibration and Regionalization, Water Resources Research, 61, e2024WR039 525, https://doi.org/10.1029/2024WR039525, 2025b.

Tanzeeba, S. and Gan, T. Y.: Potential Impact of Climate Change on the Water Availability of South Saskatchewan River Basin, Climatic Change, 112, 355–386, https://doi.org/10.1007/s10584-011-0221-7, 2012.

Taylor, K. E.: Summarizing Multiple Aspects of Model Performance in a Single Diagram, Journal of Geophysical Research: Atmospheres, 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Thornton, P. E., Running, S. W., and White, M. A.: Generating Surfaces of Daily Meteorological Variables over Large Regions of Complex Terrain, Journal of Hydrology, 190, 214–251, https://doi.org/10.1016/S0022-1694(96)03128-9, 1997.

Tyralis, H., Papacharalampous, G., and Langousis, A.: A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources, Water, 11, 910, https://doi.org/10.3390/w11050910, 2019.

Tyson, C., Longyang, Q., Neilson, B. T., Zeng, R., and Xu, T.: Effects of Meteorological Forcing Uncertainty on High-Resolution Snow Modeling and Streamflow Prediction in a Mountainous Karst Watershed, Journal of Hydrology, 619, 129 304, https://doi.org/10.1016/j.jhydrol.2023.129304, 2023.

Vernay, M., Lafaysse, M., Mérindol, L., Giraud, G., and Morin, S.: Ensemble Forecasting of Snowpack Conditions and Avalanche Hazard, Cold Regions Science and Technology, 120, 251–262, https://doi.org/10.1016/j.coldregions.2015.04.010, 2015.

Verseghy, D. L.: Class—A Canadian Land Surface Scheme for GCMS. I. Soil Model, International Journal of Climatology, 11, 111–133, https://doi.org/10.1002/joc.3370110202, 1991.

Vionnet, V., Fortin, V., Gaborit, E., Roy, G., Abrahamowicz, M., Gasset, N., and Pomeroy, J. W.: Assessing the Factors Governing the Ability to Predict Late-Spring Flooding in Cold-Region Mountain Basins, Hydrology and Earth System Sciences, 24, 2141–2165, https://doi.org/10.5194/hess-24-2141-2020, 2020.

Vionnet, V., Marsh, C. B., Menounos, B., Gascoin, S., Wayand, N. E., Shea, J., Mukherjee, K., and Pomeroy, J. W.: Multi-Scale Snowdrift-Permitting Modelling of Mountain Snowpack, The Cryosphere, 15, 743–769, https://doi.org/10.5194/tc-15-743-2021, 2021a.

Vionnet, V., Mortimer, C., Brady, M., Arnal, L., and Brown, R.: Canadian Historical Snow Water Equivalent Dataset (CanSWE, 1928–2020), Earth System Science Data Discussions, pp. 1–22, https://doi.org/10.5194/essd-2021-160, 2021b.

Vuyovich, C. and Daly, S. F.: The Chena River Watershed Hydrology Model, 2012.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

Wilks, D. S.: Chapter 8 - Forecast Verification, in: International Geophysics, edited by Wilks, D. S., vol. 100 of *Statistical Methods in the Atmospheric Sciences*, pp. 301–394, Academic Press, https://doi.org/10.1016/B978-0-12-385022-5.00008-7, 2011.

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., and Berne, A.: RainForest: A Random Forest Algorithm for Quantitative Precipitation Estimation over Switzerland, Atmospheric Measurement Techniques, 14, 3169–3193, https://doi.org/10.5194/amt-14-3169-2021, 2021.

Wood, A., Hopson, T., Newman, J., Brekke, L., Arnold, J., and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, Journal of Hydrometeorology, pp. 651–668, https://doi.org/10.1175/JHM-D-14-0213.1, 2016.

Wood, A. W. and Lettenmaier, D. P.: An Ensemble Approach for Attribution of Hydrologic Prediction Uncertainty, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL034648, 2008.

Wu, W., Emerton, R., Duan, Q., Wood, A., Wetterhall, F., and Robertson, D.: Ensemble Flood Forecasting: Current Status and Future Opportunities, Wiley Interdisciplinary Reviews: Water, 7, e1432, https://doi.org/10.1002/wat2.1432, 2020.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D.: A High-Accuracy Map of Global Terrain Elevations, Geophysical Research Letters, 44, 5844–5853, https://doi.org/10.1002/2017GL072874, 2017.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset, Water Resources Research, 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Yang, D., Goodison, B. E., Metcalfe, J. R., Golubev, V. S., Bates, R., Pangburn, T., and Hanson, C. L.: Accuracy of NWS 8" Standard Nonrecording Precipitation Gauge: Results and Application of WMO Intercomparison, 1998.

Zhao, W. and Li, A.: A Review on Land Surface Processes Modelling over Complex Terrain, Advances in Meteorology, 2015, 607 181, https://doi.org/10.1155/2015/607181, 2015.