BUILDING TRUST IN SNOWPACK MODELS: EVALUATION GUIDELINES FOR RESEARCHERS AND FORECASTERS

Simon Horton[1*1], Florian Herla[1], Alec van Herwijnen[2], Francis Meloche[2,3], Travis J. Morrison[4], Louis Quéno[2], Vincent Vionnet[5]

[1]*Simon Fraser University, Burnaby, Canada*
[2]*WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland*
[3]*Institute for Geotechnical Engineering, ETH Zürich, Switzerland*
[4]*Utah Avalanche Center, Salt Lake City, USA*
[5]*Environment and Climate Change Canada, Montreal, Canada*

ABSTRACT: Physically based snowpack models are increasingly integrated into operational avalanche hazard forecasting, yet fragmented evaluation practices reduce confidence in their outputs. Mismatched spatial and temporal scales, heterogeneous datasets, inconsistent terminology, and custom software pipelines limit reproducibility and prevent groups from transferring findings. To address these challenges, we introduce a structured evaluation framework for snowpack models in support of avalanche hazard forecasting. The framework organizes evaluation design across four core dimensions: *Context* (why and where an evaluation is done), *Model* (how the model chain is configured), *Data* (which observations are used), and *Analysis* (how model–data comparisons are performed). Applying this framework highlights the critical need to separate physical measurement agreement (*Quality*) from practical forecasting utility (*Value*). Effective evaluation requires assessing each stage of the model chain sequentially, aligning spatial and temporal scales, and reporting performance for specific scenarios rather than aggregate summaries. We translate these insights into actionable guidance for researchers, software developers, and forecasters to build trust and advance model-based decision support.

KEYWORDS: snowpack model; model evaluation; avalanche forecasting; decision support

1. INTRODUCTION

Physically based snowpack models, such as SNOWPACK and Crocus, simulate complex layered stratigraphy from atmospheric boundary conditions (Lehning et al. 1999; Lafaysse et al. 2026). By resolving mass and energy balances, snow metamorphism, and mechanical stability, these models provide forecasters with a continuous account of snowpack evolution that field observations alone cannot replicate. Operational model chains now span spatial domains from individual weather plots to regional forecasting grids (Morin et al. 2020). As these systems assume an expanding role in operational hazard assessment, rigorous model evaluation shifts from a retrospective academic exercise to a core operational requirement.

Establishing operational trust requires evaluating two distinct objectives, following Murphy's typology of forecast goodness (Murphy 1993). *Quality* measures the correspondence between model outputs and matching physical observations, whereas *Value* measures the incremental benefit of those outputs to operational decision-making. Evaluation literature routinely conflates them: statistical agreement with point-scale observations provides direct evidence of *Quality*, but cannot prove operational *Value* (Oreskes et al. 1994; Merz et al. 2024). Blurring these goals during study design obscures operational uncertainty, limits reproducibility, and hinders knowledge transfer between research groups and forecasting centers (Morin et al. 2020).

Morin et al. (2020) highlighted these fragmented evaluation practices and called for coordinated evaluation benchmarks—a mandate this paper addresses as a collective community effort building on consensus from the AvaCollabra working group (AvaCollabra 2025). We synthesize published literature and propose a structured classification framework for snowpack model evaluation in avalanche hazard forecasting. Finally, we translate these insights into actionable guidance for researchers, software developers, and operational forecasters.

2. EVALUATION FRAMEWORK AND LITERATURE SYNTHESIS

We organize our evaluation framework around four core dimensions: *Context*, *Model*, *Data*, and *Analysis*. Robust model evaluation requires conceptual alignment across all four dimensions to ensure

* *Corresponding author address:*
Simon Horton, Simon Fraser University,
8888 University Drive, Burnaby, BC V5A 1S6, Canada;
email: shorton@avalanche.ca

Table 1: Classification framework for describing snowpack model evaluation by dimensions, categories, and elements.

Dimension	Category	Elements
Context	*Purpose*	*Quality*, *Value*
	Application	*Fundamental Research*, *Applied Research*, *Operational*
	Spatial Scale	*Site-Specific*, *Regional*
	Temporal Scale	*Hindcast*, *Nowcast*, *Forecast*
Model	*Stage*	*Pre-Processing*, *Snowpack Model*, *Post-Processing*
	Weather Inputs	*Automated Stations*, *Numerical Weather Prediction*, *Analysis*, *Reanalysis*, *Hybrid*
	Spatial Geometry	*Station-Based*, *Topographic Classes*, *Gridded*
	Temporal Resolution	*Sub-Daily*, *Daily*, *Event-Based*
	Assimilation	*Open-Loop*, *Bulk-State*, *Internal-State*
Data	*Data Type*	*Weather*, *Snowpack*, *Avalanche*, *Hazard Assessment*, *Operational Decisions*
	Data Provenance	*Automated*, *Human Observation*, *Synthesis*
	Spatial Footprint	*Point*, *Slope*, *Regional*
	Temporal Sampling	*Continuous*, *Periodic*, *Event-Triggered*
Analysis	*Strategy*	*Verification*, *Calibration*, *Diagnostic*, *Validation*, *Value Assessment*

meaningful interpretation of results. The framework shares core physical evaluation principles with snow hydrology, but adds additional requirements of avalanche forecasting with a strong focus on operational value.

2.1. Context

The *Context* dimension establishes *why* and *where* an evaluation is conducted, anchoring its scope across four categories: *Purpose*, *Application*, *Spatial Scale*, and *Temporal Scale*.

Purpose explicitly distinguishes between *Quality* (physical measurement agreement) and *Value* (decision-support utility) (Murphy 1993). Historical literature is heavily dominated by *Quality* studies, whereas *Value* evaluations are less common and often conducted internally by forecasting services (Morin et al. 2020; AvaCollabra 2025). Studies can pursue both objectives, provided they are declared separately and supported by distinct evidence. A recurring pitfall is conflating these claims—treating physical accuracy as proof of decision-support skill.

Application categorizes the study environment into *Fundamental Research*, *Applied Research*, or *Operational*. Fundamental research typically uses curated datasets from instrumented reference sites (e.g., Col de Porte, Weissfluhjoch) to refine process physics; however, these setups risk overfitting when transferred across snow climates (Lejeune et al. 2019; Pérez-Guillén et al. 2026). Applied research tests full model chains on retrospective datasets or testing environments. Operational implementation embeds model chains into live workflows, where execution runtime, data availability, and operational factors constrain evaluation design.

Spatial Scale specifies the spatial domain of the forecasting context, categorized as *Site-Specific* (individual slopes/paths) or *Regional* (many slopes/paths) (European Avalanche Warning Services (EAWS) 2023). While fundamental research operates at site-specific scales where physical coherence is easily maintained, operational forecasting targets regional scales to assess broader hazard patterns. Both scales are difficult to fully characterize with 1D snow profile simulations.

Temporal Scale specifies the decision window and lead time, categorized as *Hindcast*, *Nowcast*, or *Forecast*. While research often operates on retrospective hindcast scales where rich evaluation datasets exist, operational forecasting targets nowcast and forecast scales where real-time observations are sparse or non-existent.

2.2. Model

The *Model* dimension describes how the simulation system is configured, tracking information flow and error propagation across five categories: *Stage*, *Weather Inputs*, *Spatial Geometry*, *Temporal Resolution*, and *Assimilation*.

Stage describes sequential steps: *Pre-Processing*

of weather forcings and initial snow conditions; the *Snowpack Model* that evolves a layered snowpack (e.g., SNOWPACK or Crocus); and *Post-Processing* that derives stability indices, avalanche problems, or danger ratings (Giraud 1992; Mayer et al. 2022; Reuter et al. 2022). Because uncertainties propagate downstream, interpreting outputs at later stages requires an explicit accounting of upstream errors. Evaluating bulk mass-balance properties—such as total snow depth or snow water equivalent—serves as a common language with snow hydrology to screen weather forcings before evaluating avalanche-specific stratigraphy, stability indices, or post-processed hazard outputs (Raleigh et al. 2015; Viallon-Galinier et al. 2020).

Weather Inputs identify atmospheric forcing sources: *Automated Stations*, *Numerical Weather Prediction* models, real-time meteorological *Analysis* systems (e.g., SAFRAN, CaPA), retrospective *Reanalysis* products (e.g., ERA5), or *Hybrid* blends. Weather forcing is typically the primary source of simulation error (Raleigh et al. 2015; Günther et al. 2019). Automated station inputs suffer from instrumental artifacts (wind undercatch, riming) and extrapolate poorly across complex terrain, whereas gridded NWP and analyses carry precipitation biases and smooth over local topography (Vionnet et al. 2016; Horton and Haegeli 2022).

Spatial Geometry captures how 1D snow columns are configured across terrain: *Station-Based* point profiles, *Topographic Classes* (e.g., elevation/aspect bands), or *Gridded* domains (Morin et al. 2020). *Spatial Geometry* should match the contextual *Spatial Scale* and data *Spatial Footprint* to avoid scaling errors, which often requires downscaling or upscaling weather inputs.

Temporal Resolution specifies the model output time step: *Sub-Daily*, *Daily*, or *Event-Based*. Resolution should match contextual *Temporal Scale* and data *Temporal Sampling*. For example, sub-daily models cannot be fairly evaluated with periodic *Human Observation* datasets alone.

Assimilation describes whether and how a system incorporates observations throughout a season, spanning *Open-Loop* runs (unadjusted weather forcing inputs only), *Bulk-State* assimilation (adjusting forcing or snow height), and *Internal-State* assimilation (updating layer stratigraphy) (Magnusson et al. 2017). Because snowpack models maintain seasonal memory, assimilation routines carry a risk of erasing history, over-emphasizing single observations, or creating unphysical artifacts (Viallon-Galinier et al. 2020).

2.3. Data

The *Data* dimension characterizes evaluation datasets across four categories: *Data Type*, *Data Provenance*, *Spatial Footprint*, and *Temporal Sampling*.

Data Type categorizes observations into physical measurements (*Weather*, *Snowpack*, *Avalanche*) and operational datasets (*Hazard Assessment*, *Operational Decisions*). Data type dictates which stage of the model chain can be evaluated: weather and bulk snowpack evaluate pre-processing; snow profiles evaluate the snowpack model; and avalanche activity or hazard assessments evaluate post-processing routines (Morin et al. 2020). Combining these complementary data types enables triangulation across the full model chain.

Data Provenance describes baseline structural uncertainty of a dateset: *Automated* instrument data, *Human Observation* manually recorded in the field, and *Synthesis* datasets where humans integrate broader evidence such as regional danger ratings (Techel and Schweizer 2017; Herla et al. 2025). Automated sensing provides objective metrics with defined footprints but are vulnerable to sensor artifacts. Human observations capture rich interpretations but suffer from variable practices, experience, and discontinuous coverage. Synthesis datasets integrate complex evidence across broad scales, but embed subjectivities that prevent them from serving as ground truth.

Spatial Footprint defines the physical area sampled by an observation: *Point*, *Slope*, or *Regional*. Transforming observational data to fit model geometry using aggregation or downscaling introduces spatial representativeness uncertainty that can overwhelm internal simulation errors (Rutter et al. 2014; Horton and Haegeli 2022).

Temporal Sampling describes the observation frequency: *Continuous* automated records, *Periodic* manual sampling, or *Event-Triggered* observations. Transforming temporal sampling to fit model temporal resolution or aggregating to coarser temporal scales introduces additional scaling uncertainty.

2.4. Analysis

The *Analysis* dimension characterizes the techniques used to compare models and data under a single category, *Strategy*. Five strategies form a complementary spectrum—from code integrity to

Table 2: Potential misalignments between framework dimensions.

Dimensions	Potential Misalignment
Context & *Model*	Applying a snowpack model calibrated for site-specific research directly to a regional scale without local testing
Context & *Data*	Supporting regional *Value* claims using exclusively data with point- or slope-scale footprints
Context & *Analysis*	Justifying *Value* skill using only aggregate seasonal metrics, without *Value Assessment* or event-based case studies at relevant temporal scales
Model & *Data*	Mismatching model geometry or resolution with data footprint and sampling, or transforming data without quantifying scaling uncertainties
Model & *Analysis*	Evaluating an end-to-end model chain with a single summary score, masking compensating errors across stages
Data & *Analysis*	Inflating reported performance by using data that is not fully independent of calibration or assimilation datasets

decision-support skill—aligned with common standards for verification and validation (Oreskes et al. 1994; Thacker et al. 2004; National Aeronautics and Space Administration 2026). These strategies are analytical steps that can be combined rather than mutually exclusive study types.

Verification checks internal numerical and logical consistency against closed computational criteria, establishing that algorithms, conservation laws, and software configurations function as intended (Thacker et al. 2004; National Aeronautics and Space Administration 2026). Verification does not establish physical truth for open environmental systems; rather, it proves code correctness, mass/energy balance integrity, and configuration sanity (Oreskes et al. 1994).

Calibration systematically optimizes internal physical parameters or statistical post-processing routines. Parameter sets tuned at one reference site often lose skill when transferred across snow climates (Pérez-Guillén et al. 2026). Calibration frameworks should prioritize sample efficiency—extracting maximal information from limited observations—to ensure parameters remain physically stable without overfitting.

Diagnostic approaches interrogate system mechanics, parameter sensitivities, error propagation, and relative model behavior without requiring observational ground truth (National Aeronautics and Space Administration 2026). Sensitivity analyses identify which weather inputs or physical parameters exert dominant control over outputs (Raleigh et al. 2015; Günther et al. 2019), while ensemble uncertainty quantification measures predictive spread. Model inter-comparisons benchmark different models under identical forcing, demonstrating that configuration choices often generate as much variance as core snow physics (Menard et al. 2021).

Validation measures predictive accuracy against independent observation datasets to assess empirical adequacy for a declared purpose (*Quality* or *Value*), acknowledging that open environmental systems cannot be formally verified as absolute physical truth (Oreskes et al. 1994). Validation designs should enforce strict data independence: observations should not have been used to tune parameters, train post-processors, or update assimilated state, and should be held out across entire winter seasons or geographic regions rather than randomly split days (Hendrikx et al. 2014). Crucially, reporting should avoid the “averaging trap”: good average performance can hide poor performance during the high-consequence situations that matter most for avalanche forecasting (Herla et al. 2025; Pérez-Guillén et al. 2026).

Value Assessment evaluates human–system interaction, workflow integration, and decision-support efficacy when automated guidance is embedded in forecasting operations (Morin et al. 2020). Where *Validation* primarily tests correspondence with observations (*Quality*), *Value Assessment* tests operational decision-support skill by quantifying forecaster agreement, decision efficiency, cognitive workflow, and interface usability under live constraints (Herla et al. 2025).

2.5. Framework Misalignments

A recurring vulnerability in snowpack model evaluation is unacknowledged misalignment across framework dimensions (Table 2). Conceptual alignment along spatial and temporal axes forms the structural bridge: operational *Spatial Scale* and *Temporal Scale* should dictate the model’s *Spatial Geometry* and *Temporal Resolution*, which should match the *Spatial Footprint* and *Temporal*

Table 3: Design considerations for snowpack model evaluation. Core items are typical initial priorities and refinement items usually follow once gaps are known.

Dimension	Evaluation Practice	Design Step
Context	State *Quality* vs. *Value* purpose explicitly	Core
	Define relevant *Spatial Scale* and *Temporal Scale* for the application	Core
Model	Align *Spatial Geometry* and *Temporal Resolution* with context scales	Core
	Save intermediate outputs to evaluate each *Stage* separately	Core
	If assimilating observations, maintain an uncorrected *Open-Loop* baseline to track impacts	Refinement
	Verify numerical and configuration consistency for local application	Refinement
Data	Enforce strict data independence between calibration, assimilation, and held-out evaluation	Core
	Acknowledge limitations of *Provenance*, *Spatial Footprint*, *Temporal Sampling*, and overall data representativeness	Core
	Triangulate evaluations across multiple *Data Types*	Refinement
Analysis	Check spatial/temporal alignment before interpreting metrics; avoid seasonal aggregates that hide failures	Core
	Stratify results by avalanche problem type, terrain classes, and climate regime	Core
	Benchmark model performance against simple baselines	Refinement
	Combine quantitative *Validation* with *Verification*, *Diagnostic* tests, case studies, or *Value Assessment*	Refinement

Sampling of the evaluation dataset. Divergence between these dimensions yields ambiguous or misleading claims of model skill. For example, using point-scale weather-station runs to justify regional utility, or using seasonal aggregate scores to claim *Value* for daily forecasting, masks operational failures.

3. RECOMMENDATIONS

Translating our framework into practice requires actionable guidance across the modelling ecosystem. Evaluation is situational and inherently iterative; it cannot be reduced to a rigid, prescriptive checklist. In practice, studies rarely start with a blank slate—researchers and forecasting services typically begin with whatever datasets and model outputs are already on hand. Evaluation should start by learning where those evidence streams fall short for the question being asked, and then refine their model setups, data collection, and analytical methods in subsequent cycles. This iterative loop naturally traces a path through our four framework dimensions, mapping directly to the core considerations and refinement practices outlined in Table 3.

Every evaluation cycle begins with *Context*: define the core question upfront and distinguish *Quality* from *Value* at relevant spatial and temporal scales. When setting up the *Model*, align *Spatial Geometry* and *Temporal Resolution* to that operational decision scale, and save intermediate outputs so you can evaluate pipeline stages separately. For *Data*, compare like with like in space and time, explicitly accounting for representativeness and data-independence gaps. Finally, choose an *Analysis* strategy that addresses scale alignment and stratifies results by relevant avalanche problem types and terrain classes to avoid aggregation traps that hide critical errors. Table 3 organizes these considerations into *Core* priorities for initial cycles and *Refinement* steps as systems mature.

3.1. For Researchers

Study design begins by declaring whether an evaluation tests physical measurement agreement (*Quality*) or practical forecasting utility (*Value*), choosing evaluation metrics that directly match that purpose (Murphy 1993; Morin et al. 2020). Spatial and temporal alignment should be confirmed early in design. Before evaluating detailed stratigraphy or stability metrics, researchers should perform sequential error screening: weather observations to evaluate meteorological forcing, and bulk snow variables or changes in new snow height to evaluate the combined forcing and physical *Snowpack Model* mass balance

(Raleigh et al. 2015; Richter et al. 2020).

When executing model testing or parameter tuning, *Analysis* strategies should enforce strict data independence across space and time: held-out evaluation data should remain unused in *Calibration*, machine-learning training, and *Assimilation* updates. Designs that measure predictive skill should test models on held-out winter seasons or separate geographic regions rather than randomly splitting days, which leaks weather trends into the test set and artificially inflates accuracy (Hendrikx et al. 2014). For small observation datasets, researchers should use sample-efficient cross-validation techniques.

To demonstrate genuine operational value, new physical updates or machine-learning *Post-Processing* should be benchmarked against simple baselines (e.g., persistence, fixed precipitation thresholds, or human danger ratings) (Techel et al. 2025). Results should be stratified by avalanche problem type, terrain classes, and snow climates rather than summarized into aggregate scores that can hide the averaging trap (Herla et al. 2025; Pérez-Guillén et al. 2026). Finally, publishing open datasets, evaluation scripts, and standardized benchmark tests enable fair cross-model comparisons across diverse snow climates (Pérez-Guillén et al. 2026; AvaCollabra 2025).

3.2. For Software Developers

System architecture should treat end-to-end model chains as versioned products. Execution pipelines should automatically log metadata, code versions, weather forcing sources, *Spatial Geometry*, *Temporal Resolution*, and machine-learning training datasets (Morin et al. 2020; AvaCollabra 2025). For systems assimilating real-time observations, pipelines should record ingested field data and archive a parallel uncorrected *Open-Loop* baseline run (Magnusson et al. 2017). To manage storage costs, systems can archive lightweight daily summaries (key bulk mass-balance variables and profile subsets) to detect model drift and debug without storing massive raw outputs (AvaCollabra 2025).

Interface design should directly account for operational forecasting constraints. Software dashboards should avoid overwhelming forecasters with cluttered arrays of raw physical parameters (Horton et al. 2020). Instead, visual displays should organize outputs into situational templates structured around avalanche problem types, terrain classes, and regions of interest—supporting rapid pattern recognition under time pressure (Klein 1997). Furthermore, interfaces should make model predictions fully traceable by allowing forecasters to inspect input weather forcings and intermediate processing steps. Transparent provenance enables forecasters to diagnose error propagation and evaluate model outputs efficiently.

3.3. For Forecasters

Forecasters should recognize inherent differences between the numerical representation of hazard in model chains and the complex real-world decision environment. Simulated profiles and post-processed outputs should always serve as an independent line of evidence that supports, rather than replaces, field observations and expert judgment (Oreskes et al. 1994). Because ultimate decision accountability remains with human forecasters, automated guidance earns trust only when predictions can be audited rapidly against underlying data and upstream processing steps, providing decision support without requiring a time-prohibitive full manual re-analysis.

Daily forecasting practice and formal model evaluation are distinct yet complementary. With little extra effort, operational teams can maintain simple logs recording when model outputs agreed or disagreed with human assessments and field observations, organized by region, terrain class, and avalanche problem type (Herla et al. 2025; Techel et al. 2025). Logging when guidance confirmed or challenged a daily hazard assessment builds a record of operational *Value* that can later support formal evaluation and help developers improve model designs and interfaces.

Finally, field observation programs and numerical models should be designed as a joint system. As data assimilation routines mature, forecasting teams can practice model-guided field sampling—using simulation uncertainties to direct observers toward areas where the model predicts impactful uncertainty. Field teams should complement detailed snow profiles with high-volume observations, such as snow depth probing and observed instability signs. Ultimately the needs of a field observation program will heavily depend on the operatiobnal context.

4. CONCLUSIONS

Evaluating snowpack models for avalanche hazard forecasting requires moving beyond isolated, point-scale physical comparisons toward multi-stage, decision-aware workflows. Model evaluation is not a one-time audit, but an iterative pro-

cess of refining model configurations, observational datasets, and analytical strategies as research or operational gaps are identified.

Synthesizing literature and community consensus through the AvaCollabra working group highlights four fundamental imperatives for future evaluation design. First, studies should ground their design in *Context*, explicitly separating physical accuracy (*Quality*) from decision-support utility (*Value*). Second, evaluations should track information flow through each *Model* stage—isolating weather forcing errors before attributing failures to downstream snowpack models or post-processing. Third, *Data* selection should align spatial and temporal scales to maximize representativeness. Fourth, *Analysis* strategies should avoid aggregation traps that smooth over critical, high-consequence events.

Ultimately, building operational trust relies as much on daily workflow integration as on published literature. Software developers and forecasting teams should collaborate to deploy transparent visual displays, maintain routine forecaster feedback logs, and participate shared open benchmark datasets. Adopting this unified framework across research and operations provides a clear, reproducible path forward to advance automated decision support in avalanche hazard forecasting.

ACKNOWLEDGEMENTS

We thank participants of the AvaCollabra working group for the discussions that shaped this framework.

REFERENCES

AvaCollabra (2025). AvaCollabra Validation Workshop Transcript (Spring 2025). Expert workshop transcript.

European Avalanche Warning Services (EAWS) (2023). Site-specific avalanche warning: Definitions and recommendations. EAWS recommendation document. Approved by EAWS General Assembly, Davos, 2022; working group led by C. Jaedicke.

Giraud, G. (1992). “MEPRA - An Expert System for Avalanche Risk Forecasting”. In: ISSW proceedings. ISSW proceedings; grey literature per NHESS guidelines.

Günther, D., Marke, T., Essery, R., and Strasser, U. (2019). Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing Data Error on Point-Scale Energy Balance Snow Model Performance. In: Water Resour. Res. 55.4, pp. 2779–2800. https://doi.org/10.1029/2018wr023403.

Hendrikx, J., Murphy, M., and Onslow, T. (2014). Classification trees as a tool for operational avalanche forecasting on the Seward Highway, Alaska. In: Cold Reg. Sci. Technol. 97, pp. 113–120. https://doi.org/10.1016/j.coldregions.2013.08.009.

Herla, F., Haegeli, P., Horton, S., and Mair, P. (2025). A quantitative module of avalanche hazard – comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations. In: Nat. Hazards Earth Syst. Sci. 25.2, pp. 625–646. https://doi.org/10.5194/nhess-25-625-2025.

Horton, S. and Haegeli, P. (2022). Using snow depth observations to provide insight into the quality of snowpack simulations for regional-scale avalanche forecasting. In: The Cryosphere 16.8, pp. 3393–3411. https://doi.org/10.5194/tc-16-3393-2022.

Horton, S., Nowak, S., and Haegeli, P. (2020). Enhancing the operational value of snowpack models with visualization design principles. In: Nat. Hazards Earth Syst. Sci. 20.6, pp. 1557–1572. https://doi.org/10.5194/nhess-20-1557-2020.

Klein, G. (1997). “The recognition-primed decision (RPD) model: Looking back, looking forward”. In: Naturalistic Decision Making. Ed. by Zsambok, C. E. and Klein, G. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 285–292.

Lafaysse, M., Dumont, M., De Fleurian, B., Fructus, M., Nheili, R., Viallon-Galinier, L., Baron, M., Boone, A., Bouchet, A., Brondex, J., Carmagnola, C., Cluzet, B., Fourteau, K., Haddjeri, A., Hagenmuller, P., Mazzotti, G., Minvielle, M., Morin, S., Quéno, L., Roussel, L., Spandre, P., Tuzet, F., and Vionnet, V. (2026). Version 3.0.2 of the Crocus snowpack model. In: Geosci. Model Dev. 19.13, pp. 6273–6334. https://doi.org/10.5194/gmd-19-6273-2026.

Lehning, M., Bartelt, P., Brown, B., Russi, T., Stöckli, U., and Zimmerli, M. (1999). snowpack model calculations for avalanche warning based upon a new network of weather and snow stations. In: Cold Reg. Sci. Technol. 30.1-3, pp. 145–157. https://doi.org/10.1016/s0165-232x(99)00022-1.

Lejeune, Y., Dumont, M., Panel, J. M., Lafaysse, M., Lapalus, P., Le Gac, E., Lesaffre, B., and Morin, S. (2019). 57 years (1960–2017) of snow and meteorological observations from a mid-altitude mountain site (Col de Porte, France, 1325 m of altitude). In: Earth Syst. Sci. Data 11.1, pp. 71–88. https://doi.org/10.5194/essd-11-71-2019.

Magnusson, J., Winstral, A., Stordal, A. S., Essery, R., and Jonas, T. (2017). Improving physically based snow simulations by assimilating snow depths using the particle filter. In: Water Resour. Res. 53.2, pp. 1125–1143. https://doi.org/10.1002/2016wr019092.

Mayer, S., Herwijnen, A. van, Techel, F., and Schweizer, J. (2022). A random forest model to assess snow instability from simulated snow stratigraphy. In: The Cryosphere 16.11, pp. 4593–4615. https://doi.org/10.5194/tc-16-4593-2022.

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. (2021). Scientific and Human Errors in a Snow Model Intercomparison. In: Bull. Amer. Meteor. Soc. 102.1, E61–E79. https://doi.org/10.1175/bams-d-19-0329.1.

Merz, B., Blöschl, G., Jüpner, R., Kreibich, H., Schröter, K., and Vorogushyn, S. (2024). Invited perspectives: safeguarding the usability and credibility of flood hazard and risk assessments. In: Nat. Hazards Earth Syst. Sci. 24.11, pp. 4015–4030. https://doi.org/10.5194/nhess-24-4015-2024.

Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., Gobiet, A., Hagenmuller, P., Lafaysse, M., Ližar, M., Mitterer, C., Monti, F., Müller, K., Olefs, M., Snook, J. S., Herwijnen, A. van, and Vionnet, V. (2020). Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implemen-

tations and prospects for the future. In: Cold Reg. Sci. Technol. 170, p. 102910. https://doi.org/10.1016/j.coldregions.2019.102910.

Murphy, A. H. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. In: Weather and Forecasting 8.2, pp. 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

National Aeronautics and Space Administration (Feb. 2026). NASA Handbook for Models and Simulations: An Implementation Guide for NASA-STD-7009B. NASA-HDBK-7009B; document date 2026-02-03. Office of the NASA Chief Engineer.

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. In: Science 263.5147, pp. 641–646. https://doi.org/10.1126/science.263.5147.641.

Pérez-Guillén, C., Bacardit, M., Hendrick, M., and Monti, F. (2026). From the Swiss Alps to the Pyrenees: Evaluating the transferability of machine learning models for avalanche forecasting. In: Cold Reg. Sci. Technol. 246, p. 104884. https://doi.org/10.1016/j.coldregions.2026.104884.

Raleigh, M. S., Lundquist, J., and Clark, M. P. (2015). Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. In: Hydrol. Earth Syst. Sci. 19.7, pp. 3153–3179. https://doi.org/10.5194/hess-19-3153-2015.

Reuter, B., Viallon-Galinier, L., Horton, S., Herwijnen, A. van, Mayer, S., Hagenmuller, P., and Morin, S. (2022). Characterizing snow instability with avalanche problem types derived from snow cover simulations. In: Cold Reg. Sci. Technol. 194. Published 2022; commonly cited as Reuter et al. 2022., p. 103462. https://doi.org/10.1016/j.coldregions.2021.103462.

Richter, B., Herwijnen, A. van, Rotach, M. W., and Schweizer, J. (2020). Sensitivity of modeled snow stability data to meteorological input uncertainty. In: Nat. Hazards Earth Syst. Sci. 20.11, pp. 2873–2888. https://doi.org/10.5194/nhess-20-2873-2020.

Rutter, N., Sandells, M., Derksen, C., Toose, P., Royer, A., Montpetit, B., Langlois, A., Lemmetyinen, J., and Pulliainen, J. (2014). Snow stratigraphic heterogeneity within ground-based passive microwave radiometer footprints: Implications for emission modeling. In: J. Geophys. Res. Earth Surf. 119.3, pp. 550–565. https://doi.org/10.1002/2013jf003017.

Techel, F., Purves, R. S., Mayer, S., Schmudlach, G., and Winkler, K. (2025). Can model-based avalanche forecasts match the discriminatory skill of human danger-level forecasts? A comparison from Switzerland. In: Nat. Hazards Earth Syst. Sci. 25.9, pp. 3333–3353. https://doi.org/10.5194/nhess-25-3333-2025.

Techel, F. and Schweizer, J. (2017). On using local avalanche danger level estimates for regional forecast verification. In: Cold Reg. Sci. Technol. 144, pp. 52–62. https://doi.org/10.1016/j.coldregions.2017.07.012.

Thacker, B. H., Doebling, S. W., Hemez, F. M., Anderson, M. C., Pepin, J. E., and Rodriguez, E. A. (Oct. 2004). Concepts of Model Verification and Validation. Tech. rep. LA-14167-MS. Los Alamos, NM: Los Alamos National Laboratory. https://doi.org/10.2172/835920.

Viallon-Galinier, L., Hagenmuller, P., and Lafaysse, M. (2020). Forcing and evaluating detailed snow cover models with stratigraphy observations. In: Cold Reg. Sci. Technol. 180, p. 103163. https://doi.org/10.1016/j.coldregions.2020.103163.

Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Quéno, L., Seity, Y., and Bazile, E. (2016). Numerical weather forecasts at kilometer scale in the French Alps: Evaluation and application for snowpack modeling. In: J. Hydrometeor. 17.10, pp. 2591–2614. https://doi.org/10.1175/JHM-D-15-0241.1.