ASSESSING PUBLIC BACKCOUNTRY AVALANCHE FORECASTS

Ethan Davis[1*], Zachary Miller[1], Scott Savage[1], Dallas Glass[2], Spencer Logan[3], Ethan Greene[3], Blase Reardon[4], Christine Donnelly[4], and Andrew Schauer[5]

[1] *Sawtooth Avalanche Center, Ketchum, ID, USA*
[2] *Northwest Avalanche Center, Seattle, WA, USA*
[3] *Colorado Avalanche Information Center, Boulder, CO, USA*
[4] *Flathead Avalanche Center, Hungry Horse, MT, USA*
[5] *Chugach National Forest Avalanche Center, Girdwood, AK, USA*

ABSTRACT: Avalanche danger ratings are the cornerstone of a public backcountry avalanche forecast, and the "Tier 1" danger — the highest rating across all elevation bands — is its most significant and far-reaching component. This project investigates Tier 1 danger ratings at five North American avalanche centers via forecast assessments conducted from 2019 through 2026, using two approaches: one within 24–72 hours of the forecast date and one in the off-season, both leveraging additional observed avalanche and weather information. We ask how often Tier 1 danger changes during assessment and what factors drive those changes. Previous studies found that assessed Tier 1 danger differed from forecast Tier 1 danger in 16–25% of forecasts (Techel and Schweizer, 2017; Statham et al., 2018b; Logan and Greene, 2023; Donnelly and Reardon, 2024); across 15,576 forecast-zone-days, our data show a 15.8% equal-weighted change rate, with center-level rates spanning 3.4% to 32.2%. Among changed ratings, 73% were overforecast and 27% underforecast, and almost all moved by a single danger level. Change rates rise steadily with forecast danger and fall with forecaster confidence. Storm Slab and Wet Loose problems are the most volatile. Day 2 Tier 1 ratings change 1.7–4.2 times as often as Day 1 forecasts. Danger level, forecaster confidence, and avalanche problem type predict revision across centers. Forecast assessment can identify recurring sources of mismatch and guide improvements in Tier 1 consistency within a program, but each center's rate reflects its own assessment design as much as its forecasting, so comparing programs will require a shared protocol.

KEYWORDS: regional avalanche forecasting, forecast verification, forecast assessment, avalanche danger, forecaster confidence

1. INTRODUCTION

Avalanche danger ratings are the cornerstone of a public backcountry avalanche forecast, and the overall danger rating is the single piece with the greatest influence on backcountry decision-making (Lazar et al., 2016). The Tier 1 danger, the highest rating across a forecast's elevation bands, is its most consequential component.

Producing an accurate Tier 1 rating is difficult. Avalanche danger cannot be measured directly; it must be inferred, under operational deadlines, from an incomplete and evolving mix of weather, snowpack, and avalanche-activity information, much of it predicted rather than observed. LaChapelle (1980) described conventional forecasting as an iterative, inductive process that reduces uncertainty without eliminating it, and McClung (2000) classified avalanche predictions by how much of the input data is predicted rather than observed, ranging from a true forecast to a nowcast of current instability. The Conceptual Model of Avalanche Hazard (CMAH; Statham et al., 2018a) supplies the structure within which that judgment operates: problem type, terrain location, likelihood, and expected size. Discrepancies between a forecast and its later assessment are therefore best read not as gross errors but as disagreements on specific CMAH inputs, driven by information unavailable when the forecast was written.

Even given identical information, forecasters disagree. Lazar et al. (2016) gave 68 forecasters in the United States, Canada, and New Zealand ten identical scenarios. In nine of the ten, the most-chosen danger rating drew only 40–63% of responses. Ten of the thirteen operations with three or more respondents produced at least one scenario spanning three danger ratings. Techel et al. (2024) reported that Swiss and Norwegian forecasters drafting independently for the same task agreed on danger level 76–87% of the time but agreed far less often on the underlying factor estimates. The two findings come from different hazard frameworks, suggesting the disagreement reflects judgment rather than any one system. If two qualified forecasters working simultaneously disagree 13–24% of the time, some share of every forecast-assessment difference reflects irreducible human variability rather than genuine revision.

* Corresponding author address:
Ethan Davis, Sawtooth Avalanche Center,
206 Sun Valley Rd, Sun Valley, ID 83353, USA;
email: ethan@avalanche.org

Studies measuring forecast-assessment consistency directly report a similar band. Cagnati et al. (1998) found 24-hour reliability of 76–93%; Sharp (2014) reported 81% agreement between one-day forecasts and local nowcasts, falling to 69% at three days; and Techel and Schweizer (2017), using 9,543 local estimates from trained Swiss observers, found agreement of 71–76%, with regional forecasts too high 20% of the time versus only 4% too low. Statham et al. (2018b) compared 3,752 Canadian bulletins to real-time nowcasts, reporting 81% accuracy at 24 hours and 73% across all lead times, highest at Low danger (84%) and declining above, with errors skewed toward overforecasting by one step. In Colorado, Logan and Greene (2023) found Tier 1 agreement of 84% across 1,170 assessments. In Montana, Donnelly and Reardon (2024) found agreement above 80% at Low and Moderate danger and declining above.

Read as mismatch rather than agreement rates, this literature converges to 16–25% once Donnelly and Reardon's deliberately high-danger-weighted sample is adjusted to a full season's distribution. What it does not provide is a side-by-side measurement across independently operated programs: each North American study above examines a single center using that center's own assessment design, vocabulary, and confidence scheme, so whether the patterns each identifies are properties of avalanche forecasting generally or of one program's practice is not answerable from single-center work alone. This paper takes up that challenge across five North American centers, Centers A through E, spanning 2019–2026. We ask how often a forecast's Tier 1 rating changed between issuance and assessment and by how much; in which direction; under what circumstances (which danger levels, problem types, confidence levels, and points in the season); and what reasons forecasters cite. Assessment is not a pass/fail accuracy check but a repeatable process for locating where forecaster judgment and observed outcomes most often diverge.

2. METHODS

2.1 Study design and framing

This study treats forecast assessment as a consistency and learning signal, not a measurement against ground truth. An assessment is still an expert judgment, made after the fact with additional observations, verified weather, and updated snowpack information, but a judgment nonetheless. Every row is one forecast-zone-day: a public forecast issued by one of five centers, paired with an assessment of that same forecast made afterward.

2.2 Avalanche center data sources

The combined dataset contains 15,576 forecast-zone-day rows from five centers spanning maritime, intermountain, and continental snow climates (Table 1). Centers are identified by letter throughout, consistently across every table and figure, to emphasize cross-center patterns rather than any individual program's rates.

Table 1: Avalanche centers and record counts. Coverage: C and E, continuous in-season 2019–2026; D, continuous in-season 2021–2026; B, off-season 2018 and 2021–2023; A, a targeted set of 43 off-season dates, 2022–2024.

Center	*Zones*	*Forecast-Zone-Days*	*Share of total records*
A	3	90	0.6%
B	10	1,458	9.4%
C	10	9,380	60.2%
D	3	1,188	7.6%
E	4	3,460	22.2%
Total	**30**	**15,576**	**100%**

The centers are not five equal slices: Center C alone accounts for 60% of all rows and Center A for well under 1%. They also differ in how completely individual fields are populated. Each statistic below specifies the subset it is drawn from.

2.3 Forecast and assessment process

Each center issues a daily public forecast rating for three elevation bands (upper, middle, lower) on the five-level NAPDS (Statham et al., 2010), coded 1 = Low through 5 = Extreme. The Tier 1 rating is the maximum of the three bands and is the primary outcome examined here. The dataset includes up to three avalanche problem types, a forecaster confidence level, and a Day-2 rating. Issue timing differs. Center C publishes in the afternoon for a period that runs through the following afternoon, so its forecasts have a longer effective lead time than a morning issue. Center B moved to that format in November 2022, covering 288 of its 1,458 rows; it issued the rest in the morning.

Two assessment approaches are represented. Centers C, D, and E each produce one official in-season assessment per forecast-zone-day, made within 24–72 hours of the forecast date using information that became available in the interim. Centers A and B instead use a structured off-season design that assigns two or more

independent assessors to the same forecast day and records every assessor's judgment as its own row, with no reconciliation step (Logan and Greene, 2023; Donnelly and Reardon, 2024). Assigning multiple assessors does not by itself change what quantity is estimated: each assessment remains one independent draw from the same population rate: the rate at which an independent expert, given hindsight, concludes a forecast should be revised.

2.4 Danger, problem-type, and confidence

Forecast problem type (up to three ranked slots per row) follows the nine CMAH categories (Statham et al., 2018a): dry loose, wet loose, storm slab, wind slab, persistent slab, deep persistent slab, wet slab, glide, and cornice, plus an explicit "none" value.

Confidence methodologies vary. Centers A and B do not supply confidence. Among the three that do, Center C's data-entry form offers three qualitative options later encoded as 1, 3, and 5, while Centers D and E offer a five-point scale. We collapsed all cross-center confidence comparisons to a common three-tier scale: Low (1), Moderate (2–3), High (4–5).

2.5 Derived assessment metrics

For each elevation band, we compute a signed difference as forecast minus assessment; a positive value indicates the forecast rated danger higher than the assessment (an "overforecast"). **Tier 1 Change** is the maximum forecast band rating minus the maximum assessed band rating. A companion binary variable, **Any Change**, is coded Yes if any one of the three bands differs even where Tier 1 itself does not. It is available on a substantially larger share of rows than Tier 1 Change at every center, and the two are treated as complementary throughout.

2.6 Critical-factor taxonomy

Each center records, with varying completeness, a stated primary reason for a forecast-assessment difference, a "critical factor," in the terminology introduced by Lazar et al. (2016) and formalized into a nine-category taxonomy by Logan and Greene (2023): wind, precipitation, temperature, snowpack test results, number of avalanches, avalanche character/type, avalanche size, old snow surface, and snowpack structure.

Across the five centers, 21 distinct selections exist (the nine above plus twelve additions), and only wind, precipitation, temperature, and snowpack test results are available at every center.

One center may record *number of avalanches* where another, describing the same day, records *sensitivity to triggers*, one naming what was observed, the other what caused it. Each center's form offers only some of these options, so tallying reasons across centers compares vocabularies as much as causes. For this reason, critical-factor data is treated as a within-center signal, and Table 6 reports each center's own leading factors rather than a combined ranking.

2.7 Equal-weighting

We compute each center's change rate first, then average them, counting each center once regardless of row count, and report the pooled count-weighted figure where applicable. This matters because Center C supplies 60.2% of all rows and Center E a further 22.2%: pooling makes an "all centers" statistic closer to a two-center one. We exclude centers with fewer than ten qualifying instances for a parameter from the equal-weighted figure, and we state the contributing center set wherever it differs from all five.

3. RESULTS AND DISCUSSION

We use two complementary measures throughout: *Any Change* (any elevation-band rating differs) and *Tier 1 Change* (the overall rating differs). All summary statistics are equal-weighted per Section 2.7, with pooled figures given where meaningful.

3.1 Change rates; center-to-center variation

Table 2: Change rate by avalanche center (equal-weighted).

Center	*Any Change*	*Tier 1 Change*
A	N/A*	32.2%
B	38.6%	20.9%
C	19.0%	14.8%
D	17.9%	7.5%
E	6.8%	3.4%
All centers	**20.6%***	**15.8%**

*Center A's forecasts and assessments cover only the upper elevation band, so its Any Change value is identical to its Tier 1 Change by construction and is excluded; the all-centers Any Change figure therefore averages Centers B–E, while the Tier 1 figure averages all five.

Tier 1 ratings were revised on 15.8% of assessments, equal-weighted, just below the 16–25% mismatch range prior studies converge on (Section 1). The spread in Table 2 does not rank forecast quality. The two highest-revising centers assess off-season, where an assessor has full-season outcome information and no operational

deadline; Center A's sample also oversampled higher danger levels, where change rates are steepest everywhere.

The in-season assessment design trades that hindsight for proximity: an assessment made within 24–72 hours works from fresh context. The assessor may have been in the field on the forecast day itself.

Variation among the three in-season centers (Any Change 19.0%, 17.9%, 6.8%) tracks their danger-level and problem-type mixes as much as anything intrinsic to the centers themselves, which is why the sections that follow analyze those factors directly rather than comparing centers.

3.2 Direction and magnitude of change

When the Tier 1 rating changed, the forecast had rated danger higher than the assessment in 73.0% of those changes and lower in 27.0%. This 2.7:1 ratio matches the overforecast skew reported by Statham et al. (2018b). A Tier 1 change could be computed on 6,644 forecast-zone-days; pooled across centers, 671 of those changed. The changes were overwhelmingly single-level: 662 of the 671 shifted by exactly one danger rating (471 overforecasts, 191 underforecasts), and the nine that shifted by two were all overforecasts. No two-level underforecast occurred anywhere in the dataset.

3.3 Danger level

Change rates rise consistently with forecast danger level (Table 3).

Table 3: Tier 1 Change rate by forecast danger level (equal-weighted).

Danger Level	*Tier 1 Change*
5 – Extreme	30.0%*
4 – High	34.5%
3 – Considerable	22.2%
2 – Moderate	11.9%
1 – Low	4.6%

*Only two centers have Extreme data, both n<10; the Extreme cell is pooled rather than equal-weighted given the small, imbalanced samples.

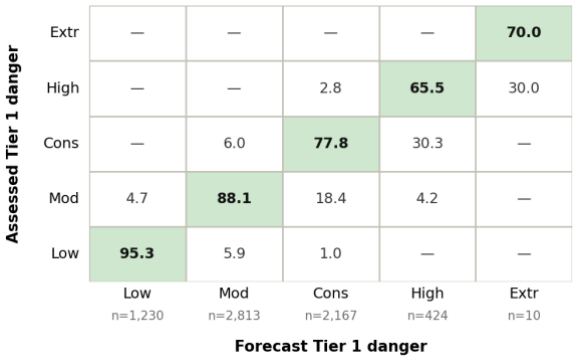

Figure 1: Contingency table of forecast versus assessed Tier 1 danger rating. Read down a column: each forecast rating sums to 100%, equal-weighted across centers with n<10 at that level excluded. The shaded diagonal is agreement. Sample sizes below each column are pooled forecast counts. Extreme is pooled rather than equal-weighted, as in Table 3.

Agreement is highest at Low and lowest at High: 95.3% at Low, 88.1% at Moderate, 77.8% at Considerable, 65.5% at High, and 70.0% at Extreme (pooled). Probability mass concentrates tightly along the diagonal at Low and Moderate, then bleeds increasingly into the adjacent below-diagonal cell as danger rises. Of all High forecasts, 30.3% were assessed Considerable and only 4.2% dropped two levels to Moderate; of all Considerable forecasts, 18.4% were assessed Moderate, 1.0% Low, and 2.8% High. Where the two judgments differ, they differ by one level and almost always downward.

Agreement declines uninterrupted from Low to High, with no rating at which it recovers, consistent with genuine difficulty discriminating among the upper danger levels rather than a problem specific to one rating.

3.4 Forecaster confidence

As forecaster confidence increases, change rates decrease. Table 4 includes Centers C, D, and E only.

Table 4: Change rate by forecaster confidence (equal-weighted, Centers C/D/E).

Confidence	*Any Change*	*Tier 1 Change*
High	8.7%	4.0%
Moderate	18.0%	8.9%
Low	18.0%	14.5%

The gradient is steeper for Tier 1 Change (4.0% to 14.5%, a 3.6-fold increase from High to Low confidence) than for Any Change (8.7% to 18.0%, roughly twofold), indicating that low confidence is disproportionately associated with revisions that

move the overall danger rating rather than a single elevation band.

Confidence is the most operationally tractable predictor identified here. Unlike danger level or problem type, it is a forecaster's own real-time self-report, available when the forecast is written rather than only in hindsight. Flagging a low-confidence forecast would be targeting a population revised 3.6 times as often as a high-confidence forecast.

3.5 Avalanche problem type

Table 5: Tier 1 Change rate by avalanche problem type.

Problem type	Equal-wtd rate	Ratio to baseline	Centers above baseline
Wet Loose	25.6%	**1.44×**	4 of 4
Glide	21.3%	1.43×*	1 of 1
Wind Slab	17.3%	1.10×	4 of 5
Persistent Slab	13.6%	0.86×	2 of 5
Storm Slab	12.5%	**1.46×**	3 of 3
Wet Slab	10.4%	0.79×	1 of 3
Deep Slab	4.5%	0.83×	0 of 2

Rates are equal-weighted across centers with at least ten qualifying forecasts of that type. A forecast qualifies for a problem type if that type was its primary forecast problem and its Tier 1 rating could be compared with the assessment; centers below ten are excluded from both the rate and the center count. Qualifying forecasts: Wet Loose 585, Glide 47, Wind Slab 1,301, Persistent Slab 2,228, Storm Slab 229, Wet Slab 84, Deep Slab 53. Baselines are matched to the contributing center set: Wet Loose 17.8%; Wind Slab and Persistent Slab 15.8%; Storm Slab 8.6%; Wet Slab 13.1%; Deep Slab 5.4%. *Glide qualifies at Center C only, so its ratio compares that center against its own overall rate (14.8%) and is not a cross-center signal. No center reached ten qualifying forecasts of Dry Loose or Cornice.

Read against each center's own baseline, two problem types stand out, and one fades. Forecasts naming Storm Slab as the primary problem were revised 1.46 times as often as a typical forecast at the same centers, and Wet Loose 1.44 times as often; neither is an artifact of one program. Storm Slab is elevated at all three centers with enough data to measure it, Wet Loose at all four. Deep Slab sits below baseline at both of its centers. Persistent Slab, by contrast, is above baseline at two of five centers and below at three, spanning 0.50× to 1.27×; its 0.86× average conceals disagreement between programs rather than a shared property, and its low pooled rate (5.7%) is mostly a composition effect.

The split has a practical shape. The types below baseline — Persistent Slab, Deep Slab, Wet Slab — are defined by structure already in the snowpack, though each hides a spread of its own: Wet Slab's 0.79× average runs from 0.40× to 1.31× across its three centers. Those at or above it — Storm Slab, Wet Loose, Wind Slab — depend on something happening within the forecast period: a storm delivering its predicted load, a warm-up producing the expected wet activity. Describing what is already there appears more reliable than anticipating what the weather has yet to deliver (Section 1). Revision direction does not separate the groups: they run 78–82% downward for every problem type with enough data to measure. Overforecasting is general, not a property of particular problems.

3.6 Critical factors

Table 6: Leading stated reasons for an Any Change (Day 1), by center (top three by each center's own rate).

Center	#1 reason	#2 reason	#3 reason
C (n=684)	Likelihood of avalanches — 35.4%	Precipitation — 28.8%	Avalanche character — 7.6%
B (n=556)	Number of avalanches — 34.2%	Precipitation — 18.0%	Snowpack structure — 15.5%
E (n=197)	Sensitivity to triggers — 27.4%	Number of avalanches — 17.8%	Avalanche distribution — 14.2%
D (n=44)	Wind — 40.9%	More stable than anticipated — 11.4%	Precipitation — 11.4%
A (n=29)	Number of avalanches — 48.3%	Temperature — 17.2%	Precipitation — 13.8%

Center C's "Likelihood of avalanches" is a CMAH composite of sensitivity to triggers and spatial distribution, which other centers report separately, so it is not comparable cell-for-cell to the more granular reasons (Section 2.6). A consistent theme emerges nonetheless: observed avalanche activity — logged as number of avalanches, sensitivity to triggers, or folded into a likelihood of avalanche composite — is the most- or second-most-cited reason at four of five centers, and precipitation appears in the top three everywhere except E. Center D is the outlier, citing wind nearly twice as often as any other factor.

Reason data of this kind is inherently noisier than the danger ratings it explains. Techel et al. (2024) found forecaster agreement consistently lower on the individual factors behind a danger judgment (51–75%) than on the danger level itself (76–87%), and that Swiss agreement would have

fallen from 87% to 54% had those factor estimates been combined mechanically rather than judged directly; decomposition compounded the variability rather than canceling it. Table 6 is therefore best read as a record of what forecasters reported as driving a revision, not as a reliable decomposition of why it occurred.

3.7 Day 2 change rates

Day 2 ratings are substantially less reliable than Day 1 forecasts. Tier 1 Change runs 1.7 to 4.2 times higher at the three centers that supplied assessed Day 2 ratings (Table 7).

Table 7: Day 1 versus Day 2 Tier 1 Change rate, by center.

Center	*Day 1*	*Day 2*	*Increase*
C	14.8%	24.8%	1.7×
D	7.5%	18.3%	2.4×
E	3.4%	14.2%	4.2×
Combined (C/D/E)	**8.6%**	**19.1%**	**2.2×**

Because only three centers supplied assessed Day 2 ratings, these figures are read against the same three centers' Day 1 rates rather than the all-center values in Table 3.

3.8 Temporal patterns

Tier 1 Change rate varies through the season, rising from November into a January–February peak before easing modestly through spring (Figure 2).

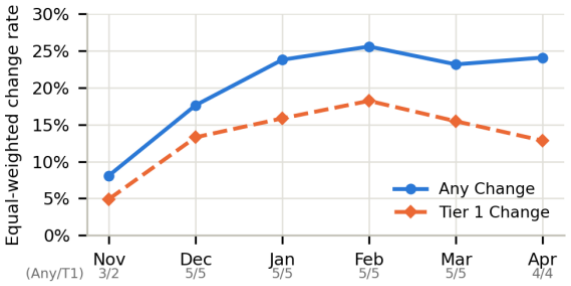

Figure 2: Equal-weighted Any Change and Tier 1 Change rate by month. The values below each month give the number of centers contributing to that month's Any Change and Tier 1 Change values, respectively.

Both series trace the same shape: a steep early-season climb (Any Change from roughly 8% in November to 24% in January; Tier 1 Change from roughly 5% to 16%), a February peak (25.6% and 18.2%), and a mild pullback into March before a slight rebound in Any Change — but not Tier 1 Change — in April. The gap between the series widens from roughly 3 points in November to 7–11 points from January onward, meaning an increasing share of late-season changes are band-level adjustments that leave the overall rating untouched.

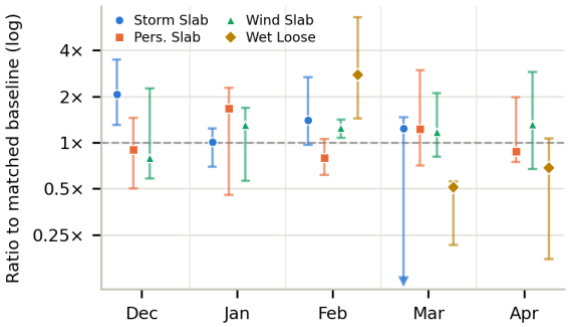

Figure 3: Which avalanche problems most often accompanied a Tier 1 revision, month by month. For each problem, its own Tier 1 Change rate that month divided by the overall rate at the same centers **(Section 2.7)**: 2.0× means forecasts naming that problem were revised twice as often as a typical forecast that month. Markers show the equal-weighted center value and whiskers span the min–max range across qualifying centers. The ratio axis is logarithmic, so equal distances above and below the 1.0× line are equal proportional departures; a downward arrow marks a whisker running to a center that revised no forecasts of that type.

Figure 3 shows avalanche problems most often accompanying a Tier 1 revision, month by month. Wet Loose's February ratio of 2.76× is the largest in the figure and carries the widest band (1.4× to 6.6× across two qualifying centers). Storm Slab swings most across the season (2.07× in December, 1.01× in January, 1.40× in February, 1.23× in March). Wind Slab and Persistent Slab, the two types with the largest samples, stay nearer baseline (Wind Slab within 0.80×–1.31× in every month), though Persistent Slab reaches 1.67× in January on three centers that disagree sharply among themselves (0.46× to 2.28×).

3.9 Limitations

Three limitations bound these results. First, the five centers' records are not equally complete, and the incompleteness is structural rather than random: every statistic is scoped to whichever subset of rows can support it, and that subset is rarely a random sample of a center's forecast-zone-days. The two assessment designs compound this, since a same-day assessor and a months-later assessor do not work from the same evidence.

Second, the five centers' reason taxonomies are not interchangeable: they overlap conceptually rather than merely differing in wording, and each form offers only some of the available labels, so

the reasons in Table 6 are comparable within a center but not across them (Section 2.6).

Third and most fundamentally, the outcome measure is agreement between two expert judgments, not accuracy against an independently verifiable danger level. Both are inferences from an incomplete observation set (LaChapelle, 1980), the assessment simply made at a more informationally complete point (McClung, 2000). No single assessment should be read as ground truth. The scale's categories are broad, and the judgment behind them is subjective, so a rating is better thought of as defensible or indefensible than as correct or incorrect: as Spencer has put it, "there is not a right answer, but there are definitely wrong ones." Ebert and Milne (2022) add a scope constraint that applies directly here: the larger the area a forecast covers, the less closely its verification can speak to conditions at any particular location.

4. CONCLUSIONS

Formal forecast assessment is not yet standard practice. Most North American avalanche centers do not reassess their forecasts, and the five here did so under different designs. Across 15,576 forecast-zone-days, Tier 1 ratings were revised on 15.8% of assessments, equal-weighted, with center-level rates spanning more than a ninefold range and downward corrections outnumbering upward ones by roughly 2.7:1. None of this shows forecasters were wrong: an assessment is a second expert judgment, not a verified outcome, and the inter-rater literature (Lazar et al., 2016; Techel et al., 2024) suggests a meaningful share of any gap reflects irreducible human variability. It shows that the differences are common enough to matter and patterned enough to act on.

Forecaster confidence is the most compelling of those patterns, and the most immediately usable. Tier 1 change rate falls from 14.5% at low confidence to 4.0% at high, a 3.6-fold difference. Low confidence is not merely a statement about the forecaster: it predicts that the same forecast, reassessed with better information, will more often carry a different danger rating. It is a real-time indicator of uncertainty in the rating, recorded before the forecast is published. Latosuo et al. (2024) found that stating uncertainty explicitly raises readers' awareness of it, supports more conservative decisions, and increases rather than erodes trust, though their readers also rated more uncertain forecasts as less useful for trip planning. However, that cost is bounded: low confidence is assigned on roughly one forecast in ten, so a confidence-triggered uncertainty statement would be the exception rather than a standing caveat. A center already recording confidence has a structured, repeatable basis for deciding when uncertainty language belongs in the product, instead of adding it at the forecaster's discretion.

Other patterns uncovered in this analysis can inform a forecast while it is still being written. Change rate climbs steadily with forecast danger, from 4.6% at Low to 34.5% at High, and problem type adds an additional signal: forecasts resting on a pending weather event are revised more often than those hinging on existing snowpack structure. The practical version is narrow. A forecast issued at high danger with low confidence that hinges on a weather forecast is the one worth a second look. Day 2 forecasts deserve the same scrutiny: their Tier 1 ratings change 1.7 to 4.2 times as often as Day 1.

Comparing programs will require agreement the community does not yet have. Two centers here record multiple independent assessments with no reconciliation step while three record a single official one; confidence is captured on incompatible scales; and the critical-factor taxonomy has fragmented into twenty-one selections across five centers. A shared protocol needs a common confidence rating, an agreed assessor design, and a unified taxonomy. Readopting the nine categories Lazar et al. (2016) and Logan and Greene (2023) set out will not settle it, because the twelve additions since were deliberate, each filling a gap a team judged important. It has to be built from those additions, collectively.

Assessment does not make avalanche danger measurable, but it turns an unmeasurable judgment into a repeatable, low-cost signal. This dataset shows that signal is not noise. Recording it consistently, reviewing it deliberately under the conditions identified above, and adopting it at centers that do not yet assess at all is a realistic next step toward more consistent Tier 1 danger forecasting.

ACKNOWLEDGEMENTS

We thank the many forecasters who contributed their time, notes, and critical judgment to building this assessment dataset.

REFERENCES

Cagnati, A., Valt, M., Soratroi, G., Gavaldà, J., and Sellés, C. G., 1998. A field method for avalanche danger-level verification, Ann. Glaciol., 26, 343–346, https://doi.org/10.3189/1998AoG26-1-343-346.

Donnelly, C. and Reardon, B., 2024. How consistent are we? Hindcasting FAC products 2023 and 2024, presented at the Northern Rockies Snow and Avalanche Workshop, Whitefish, MT, USA, 9 November 2024.

Ebert, P. A. and Milne, P., 2022. Methodological and conceptual challenges in rare and severe event forecast verification, Nat. Hazards Earth Syst. Sci., 22, 539–557, https://doi.org/10.5194/nhess-22-539-2022.

LaChapelle, E. R., 1980. The fundamental processes in conventional avalanche forecasting, J. Glaciol., 26, 75–84, https://doi.org/10.3189/S0022143000010601.

Latosuo, E., Haegeli, P., Demuth, J., and Greene, E., 2024. Users' awareness and response to uncertainty information in public avalanche forecasts, in: Proceedings of the International Snow Science Workshop, Tromsø, Norway, 1618–1625.

Lazar, B., Trautman, S., Cooperstein, M., Greene, E., and Birkeland, K., 2016. North American avalanche danger scale: do backcountry forecasters apply it consistently?, in: Proceedings of the International Snow Science Workshop, Breckenridge, CO, USA, 457–465.

Logan, S. and Greene, E., 2023. Assessing the Colorado Avalanche Information Center's backcountry avalanche forecasts, in: Proceedings of the International Snow Science Workshop, Bend, OR, USA, 518–525.

McClung, D. M., 2000. Predictions in avalanche forecasting, Ann. Glaciol., 31, 377–381, https://doi.org/10.3189/172756400781820507.

Sharp, E., 2014. Avalanche forecast verification through a comparison of local nowcasts with regional forecasts, in: Proceedings of the International Snow Science Workshop, Banff, AB, Canada, 475–480.

Statham, G., Haegeli, P., Birkeland, K., Greene, E., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J., 2010. The North American public avalanche danger scale, in: Proceedings of the International Snow Science Workshop, Lake Tahoe, CA, USA, 117–123.

Statham, G., Haegeli, P., Greene, E., Birkeland, K., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J., 2018a. A conceptual model of avalanche hazard, Nat. Hazards, 90, 663–691, https://doi.org/10.1007/s11069-017-3070-5.

Statham, G., Holeczi, S., and Shandro, B., 2018b. Consistency and accuracy of public avalanche forecasts in Western Canada, in: Proceedings of the International Snow Science Workshop, Innsbruck, Austria, 1491–1495.

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

Techel, F., Lucas, C., Müller, K., Pielmeier, C., and Morreau, M., 2024. Unreliability in expert estimates of factors determining avalanche danger and impact on danger level estimation using the matrix, in: Proceedings of the International Snow Science Workshop, Tromsø, Norway, 264–271.