Instruments and Methods

On optimal stability-test spacing for assessing snow avalanche conditions

Karl W. BIRKELAND,[1] Jordy HENDRIKX,[2] Martyn P. CLARK[3]

[1]*USDA Forest Service National Avalanche Center, PO Box 130, Bozeman, Montana 59771, USA*
E-mail: kbirkeland@fs.fed.us
[2]*National Institute of Water and Atmospheric Research Ltd, PO Box 8602, Christchurch, New Zealand*
[3]*National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80307-3000, USA*

ABSTRACT. Assessing snow stability requires a holistic approach, relying on avalanche, snowpack and weather observations. Part of this assessment utilizes stability tests, but these tests can be unreliable due in part to the spatial variability of test results. Conducting more than one test can help to mitigate this uncertainty, though it is unclear how far apart to space tests to optimize our assessments. To address this issue we analyze the probability of sampling two relatively strong test results over 25 spatial datasets collected using a variety of stability tests. Our results show that the optimal distance for spacing stability tests varies by dataset, even when taking the sampling scheme and stability-test type into account. This suggests that no clear rule currently exists for spacing stability tests. Our work further emphasizes the spatial complexity of snow stability measurements, and the need for holistic stability assessments where stability tests are only one part of a multifaceted puzzle.

INTRODUCTION

Snow avalanches are a significant hazard in mountainous areas worldwide. In the United States, avalanches kill about 30 people annually, more than the average annual death toll due to earthquakes or other mass movements (Voight and others, 1990). Determining avalanche conditions requires a holistic approach, whereby a person assesses the terrain, weather and current snowpack conditions (Fredston and Fesler, 1994; McClung and Schaerer, 2006; Tremper, 2008). Evaluating snowpack conditions can be particularly challenging. To assist in this challenge, avalanche forecasters employ snow stability tests to assess the potential for avalanching when they do not observe obvious signs of instability.

A variety of snow stability tests exist, including the compression test (Jamieson, 1999), stuffblock test (Birkeland and Johnson, 1999), quantified loaded column test (Landry and others, 2001), rutschblock test (Föhn, 1987a) and shear frame test (Perla and others, 1982). Newer tests are also becoming available, such as the extended column test (Simenhois and Birkeland, 2009) and the propagation saw test (Gauthier and Jamieson, 2008). The procedures for these (and other) tests are outlined by Greene and others (2009). All these tests provide the observer with valuable information, but there is also a great deal of uncertainty associated with test results. In fact, work shows that most tests generally have a false stability rate around 10%, meaning that on unstable slopes there is approximately a 1 in 10 chance of obtaining a stable test result (Birkeland and Chabot, 2006). This value is too high since making such an error could well result in serious injury or death. A primary reason for this false-stability rate may be the large amount of spatial variability on potential avalanche slopes (Schweizer and others, 2008).

Birkeland and Chabot (2006) suggest conducting more than one stability test on a slope to minimize the chances of incorrectly assessing an unstable slope as stable, while Schweizer and Bellaire (2009) propose conducting up to two sets of two tests 10–15 m apart, depending on the results of the first set of tests. However, neither study offers guidance for optimizing test spacing. Test spacing should insure that test results are not spatially autocorrelated, thereby minimizing the chances of obtaining two misleading test results on the same slope. Schweizer and others (2008) review several studies with varying autocorrelation lengths, and suggest, based on limited analysis, spacing tests at >10 m.

Given the nature of many snow stability spatial datasets, we need new techniques to assess optimal test spacing. The purpose of this paper is to comprehensively evaluate the probability of obtaining two stable test results. This is done by examining 25 datasets on the spatial variability of slope stability from different mountain environments around the world. Assessing slope stability requires searching for instability, so our technique quantifies the distance at which an observer is *unlikely* to obtain two 'strong' test results. We initially define a strong test as one that previous literature defines as a stable test result, and we then examine the 75th percentile of our data to better understand the spatial patterns that exist in those datasets. In essence, we are asking, 'Given a single strong stability-test result, at what distance do we minimize our chances of collecting a second strong stability-test result?' Our goal is to examine the range of these optimal distances for our datasets to provide guidance for backcountry recreationists and avalanche professionals for optimizing stability-test spacing.

METHODS

Data

Our data come from diverse sources, utilizing four different snow stability tests and a variety of spatial layouts. The support, spacing and extent (Blöschl, 1999) varies between

Table 1. The spatial datasets utilized for this paper

Datasets	Source	Stability test*	Support	Smallest spacing	Max. extent	Threshold†
			m^2	m	m	
1–10	Landry and others (2004)	QLCT	0.09	0.5	28	≥2.0
11–16	Logan and others (2007); Lutz (2009)	SF	0.025	0.5	13	≥2.0
17–21	Unpublished	SB	0.09	0.5–10	72	≥0.5 m
22–23	Unpublished	CT	0.09	10	64	≥21 taps
24–25	Campbell and Jamieson (2007)	RB	3.0	2.5	36	≥6

*QLCT: quantified loaded column test (Landry and others, 2001); SF: shear frame (Perla and others, 1982); SB: stuffblock (Birkeland and Johnson, 1999); CT: compression test (Jamieson, 1999); RB: rutschblock (Föhn, 1987a).
†Defined threshold for a 'strong' stability test.

the datasets (Table 1). Though this variability adds complications to our comparisons (Kronholm and Birkeland, 2007), we think we have enough datasets with similar spatial layouts to compare them against each other in addition to comparing them with different datasets. We provide a brief discussion of each dataset based on the stability tests used, and refer the reader to the original work for more in-depth descriptions of the data.

The first ten datasets that we analyze were collected by Landry and others (2004) utilizing the quantified loaded column test (QLCT; Landry and others, 2001). The QLCT involves manually pressing down on a 0.30 m × 0.30 m rigid plate with a gauge to assess the vertical force necessary to fracture a buried weak layer. Measurements of slope angle and slab shear stress allow the calculation of weak-layer shear strength and an associated stability index. Collected in southwest Montana, USA, each of these datasets involved five sets of ten closely spaced (0.50 m) measurements within a 30 m × 30 m area (Table 1; Fig. 1).

Our next seven datasets used the shear frame (SF) test (Perla and others, 1982; Föhn, 1987b) to quantify snow stability. These datasets were collected by Logan and others (2007) and Lutz (2009). The shear frame quantifies the shear strength, while associated measurements of slope angle and slab shear stress allow the calculation of a stability index. We also collected these data in southwest Montana. Each dataset consists of around 70 measurements of about 0.16 m × 0.16 m sampled in a 14 m × 14 m area, with a minimum distance between tests of 0.50 m (Table 1; Fig. 1).

Our next five datasets index snow stability using the stuffblock (SB) test (Birkeland and Johnson, 1999) and have not been published before. The stuffblock provides ordered data based on the height from which a nylon sack filled with 4.5 kg of snow must be dropped onto a shovel to fracture a buried weak layer in an isolated column of 0.30 m × 0.30 m. The spatial layouts of our stuffblock data vary between our five datasets. We collected two datasets concurrently with Landry and others (2004) using that spatial layout, and two datasets concurrently with Hendrikx and others (2009) using that spatial layout. Our final stuffblock dataset utilized the same slope and a similar layout to Logan and others (2007), but was collected during a different winter (Table 1; Fig. 1). Southwest Montana again served as our study area for these five datasets.

Compression tests (CTs; Jamieson, 1999) index the snow stability for our next two datasets, which we collected adjacent to New Zealand's Mount Hutt ski area, in the Eastern Coastal Range of the South Island. Compression tests are similar to stuffblock tests, with the same 0.30 m × 0.30 m support, but the load to cause weak-layer fracture is provided by a person tapping on a shovel rather than dropping a nylon sack of snow onto the shovel. Hendrikx and Birkeland (2009) compare one of these datasets with extended column test (Simenhois and Birkeland, 2009) results, but none of the three datasets has been analyzed in detail or presented in a refereed publication. The spatial layout of the data is the same as the work by Hendrikx and others (2009), with a measurement spacing of 10 m and a larger extent than the other datasets (Table 1; Fig. 1).

Our final two datasets utilize the rutschblock (RB) test (Föhn, 1987a). Developed in Switzerland, the rutschblock involves a skier progressively loading a large (2 m × 1.5 m), isolated block of snow until a buried weak layer fractures. Campbell and Jamieson (2007) collected these data in Canada's Columbia Mountains; we utilize data from their figures 6 and 9 for our analyses. The spatial layout of the data consists of regular grids (Table 1; Fig. 1).

Of our 25 datasets, 22 (88%) are at or below treeline, and 18 (72%) are not significantly affected by the wind (Table 2). The layer of interest was a persistent weak layer in 23 of the datasets (92%), while in two cases the weak layer consisted of decomposing fragments of precipitation particles. Slope elevations varied from 1900 m to almost 2700 m, while slope angles varied from 25° to 34° (Table 2). Most of the slopes (datasets 1–23) were chosen for what observers believed were reasonably consistent snowpack conditions across the slope. In other words, these slopes were selected because they appeared to be sites that could be used by an experienced observer as a test slope (Greene and others, 2009).

Data analysis

Our data analysis focuses on the following question: If a person samples one strong stability test, at what distance will that person minimize their chances of sampling a second strong test? This analysis requires defining a threshold for what constitutes a strong stability-test result. For the stability indices calculated using the shear frame and quantified loaded column tests, we chose a value of ≥2 based on Föhn (1987b), who stated that stability index values ≥1.5 suggest relatively stable conditions. Our threshold for rutschblock numbers is ≥6 (Föhn, 1987a), for stuffblock drop heights is ≥0.50 m (Birkeland and Johnson, 1999) and for compression tests is ≥21 taps (Jamieson, 1999).

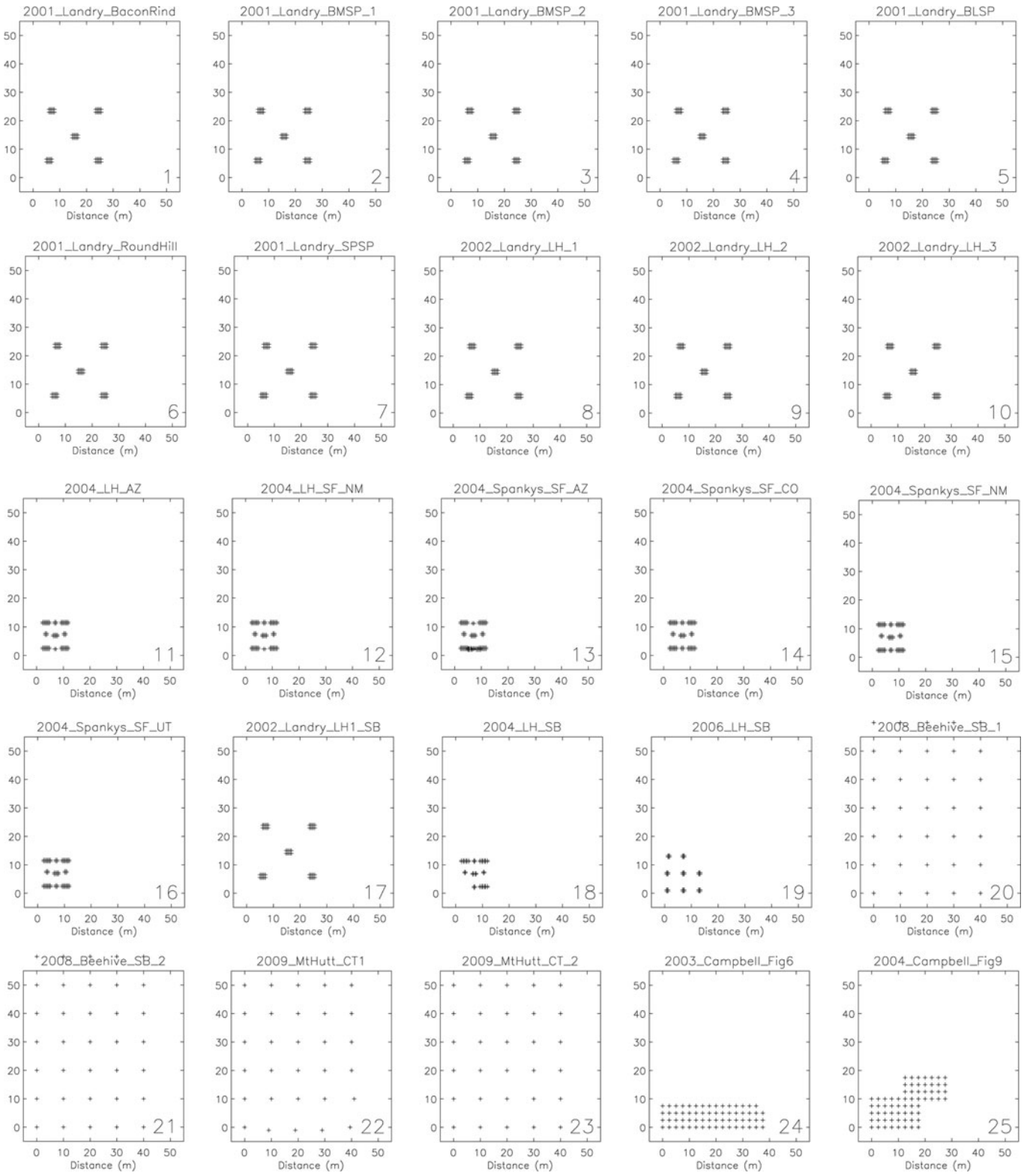

Fig. 1. The spatial layout (m) varied for our different datasets. Note that in grids 1–19 there are multiple adjacent pits.

Our analysis follows three main steps:

1. We examine the cumulative distribution function (CDF) for each spatial dataset to identify which datasets consist primarily of measurements either above or below our prescribed stability thresholds.
2. We bin the data into different distance categories, and for each distance category we compute the probability of obtaining two strong test results. This is computed as the number of data pairs in a given distance category where both points in the data pair are defined as 'stable' (i.e. above the stability threshold), divided by the total number of data pairs in that distance category. This data analysis strategy is similar to an indicator variogram that is commonly used in geostatistics (e.g. Webster and Oliver, 2001). Such indicator variograms summarize the overall spatial variability of binary data in each distance category, which is proportional to the fraction of data

Table 2. Slope characteristics associated with our datasets

Dataset(s)	Location	Altitude	Treeline?*	Aspect	Approx. slope angle	Wind-affected?	Weak layer
		m			°		
1	Montana, USA	2250	BTL	ESE	32	No	Depth hoar
2–4	Montana, USA	2320	BTL	ENE	25	Yes	Faceted crystals
5	Montana, USA	2200	BTL	ENE	25	No	Faceted crystals
6	British Columbia, Canada	2150	TL	SE	28	Yes	Surface hoar
7	Montana, USA	2240	BTL	E	26	No	Depth hoar
8–12, 17–19	Montana, USA	2340	BTL	NE	28	No	Surface hoar
13–16	Montana, USA	2670	BTL	E	27	No	Surface hoar
20–21	Montana, USA	2530	BTL	WSW	31	No	Depth hoar
22–23	Canterbury, New Zealand	2010	ATL	NW	30	Yes	Decomposing fragments
24	British Columbia, Canada	2000	ATL	ENE	28	Yes	Surface hoar
25	British Columbia, Canada	1900	TL	NE	34	No	Surface hoar

*Location of site in relation to treeline (ATL: above treeline; TL: at treeline; BTL: below treeline).

pairs where one point is stable and the other is unstable. Our approach of focusing attention on the fraction of data pairs in which both stability estimates are classified as stable is favored over more standard geostatistical methods because it directly addresses the question under investigation.

3. In some cases, using the stability thresholds defined above does not allow us to explore the spatial variations that exist in those data. For example, if all the measurements in a dataset are so strong they are above the threshold, then the probability of making two strong measurements at any distance is 1, and if all measurements are less than the threshold the probability is 0. As such, we also conducted an analysis whereby we define the threshold as the value of the 75th percentile of each specific dataset. This allows us to investigate the chances of obtaining two *relatively* strong measurements in a given dataset, and to explore more effectively the spatial relationships in each dataset.

RESULTS AND DISCUSSION

Our datasets are diverse, demonstrating a range of CDFs (Fig. 2). The CDFs also graphically demonstrate the continuous (QLCT and SF tests) and ordered (SB, CT and RB tests) nature of our different datasets. A number of our datasets (24%) represent quite stable conditions, with all values above the strong stability-test thresholds we set for that particular test (datasets 2, 8, 9, 13, 14 and 15 (Table 3)). On the other hand, a few of our datasets (16%) represent much less stable conditions, with all test results below our set thresholds (datasets 17, 18, 22 and 23 (Table 3)). We were still able to sample these less stable slopes safely because the slope angles are generally just below the threshold for avalanching. The diversity of our data allows us to investigate a wide range of conditions and tests.

Determining an optimal distance to minimize the chances of obtaining two strong stability tests is difficult for many of our datasets when we use our predetermined thresholds for a strong stability test (QLCT $\geq$2.0 m, SF $\geq$2.0 m, SB $\geq$0.50 m, CT $\geq$21 m, RB $\geq$6 m) (Fig. 3). Some datasets show a clear differentiation between various distances, such as dataset 6 where we can see that distances between tests of either 0–5 m or 15–20 m have the greatest probability of having two strong tests, while the other distances minimize that probability. However, there are a number of datasets where there is little differentiation between distances (e.g. dataset 15).

If we adjust our strong stability-test results from the set thresholds discussed above to $\geq$75th percentile for each dataset, the spatial patterns in each dataset become much more evident. This allows us to more effectively explore how to space tests to minimize the probability of sampling two *relatively* strong test results in a given dataset (Fig. 4). For most of the datasets, there are clearly certain distances (or a range of distances) which minimize that probability (Fig. 4; Table 3). Thus, an optimal sampling strategy that searches for instability with two tests on a slope will aim to conduct those tests at that distance.

Interestingly, our datasets demonstrate a range of optimal sampling distances, even when taking into account the sampling strategy and the test (Fig. 4; Table 3). For example, datasets 1–10 use the QLCT and the same basic sampling layout (Landry and others, 2004). Within these data the distance required to minimize the probability of sampling two strong tests varies from 10 to 30 m (30 m is the maximum extent of these samples). In some datasets (e.g. 2 and 3), a distance of 10–15 m will minimize the probability of sampling two strong tests. However, in dataset 9 a distance of 10–15 m maximizes this probability. We do see that close distances (<5 m) are unlikely to minimize the chances of two strong tests and, in general, longer distances tend to be better. In 8 of the 10 QLCT datasets (80%) the longest distance has the lowest, or close to the lowest, chance of two strong tests. However, in four of these cases there is also a minimum at a shorter distance, and dataset 7 actually has a spike in the probability at distances of 25–30 m (Fig. 4). Thus, for these data there appears to be no clear rule of thumb for optimal stability-test spacing.

Our next six datasets (11–16) all utilize the shear frame test and have the same layout (Logan and others, 2007; Lutz, 2009). Like the QLCT datasets, these data also demonstrate some striking variability in results (Fig. 4; Table 3). In half of these datasets the probability of sampling two strong tests is minimized at a distance of 12–14 m (datasets 13, 14 and 16), which is the maximum extent of this sampling layout. Conversely, in one case (dataset 12) we minimize the

Table 3. Summary statistics and distances which minimize the probability of sampling two strong (≥75th percentile) stability tests for each of our datasets

Dataset	Name	Test*	Min	0.25	0.50	0.75	Max	Thresh.†	Prob. ≥Th.§	*n*	Distance to minimize two strong (≥75th percentile) tests‡
											m
1	2001_Landry_BaconRind	QLCT	1.232	1.727	1.841	2.208	4.393	2.000	0.380	50	25–30, 15–20
2	2001_Landry_BMSP_1	QLCT	3.672	4.564	5.110	6.710	9.644	2.000	1.000	46	10–15, 25–30
3	2001_Landry_BMSP_2	QLCT	0.936	1.300	1.674	1.775	2.110	2.000	0.100	20	10–15
4	2001_Landry_BMSP_3	QLCT	1.952	2.386	2.987	3.648	4.900	2.000	0.978	46	15–20, 25–30
5	2001_Landry_BLSP	QLCT	1.312	1.585	1.704	1.882	2.967	2.000	0.211	19	15–30
6	2001_Landry_RoundHill	QLCT	1.035	1.535	2.036	3.683	4.369	2.000	0.514	37	10–15, 20–30
7	2001_Landry_SPSP	QLCT	0.803	1.523	1.897	2.132	2.979	2.000	0.404	47	15–20, 20–25
8	2002_Landry_LH_1	QLCT	2.021	2.389	2.498	2.646	3.607	2.000	1.000	50	25–30, 10–15
9	2002_Landry_LH_2	QLCT	2.488	2.896	3.069	3.210	4.395	2.000	1.000	48	25–30, 15–20
10	2002_Landry_LH_3	QLCT	1.731	2.108	2.370	2.642	3.311	2.000	0.880	50	15–20, 25–30
11	2004_LH_AZ	SF	1.899	2.431	2.643	3.117	4.603	2.000	0.944	72	10–12
12	2004_LH_SF_NM	SF	1.507	2.027	2.231	2.504	3.017	2.000	0.764	72	0–2
13	2004_Spankys_SF_AZ	SF	2.064	3.226	3.517	3.871	4.693	2.000	1.000	90	12–14
14	2004_Spankys_SF_CO	SF	2.085	2.687	3.103	3.362	3.985	2.000	1.000	72	12–14
15	2004_Spankys_SF_NM	SF	2.421	3.171	3.415	3.627	4.529	2.000	1.000	72	6–8, 12–14
16	2004_Spankys_SF_UT	SF	1.508	2.248	2.518	2.788	3.560	2.000	0.889	72	12–14
17	2002_Landry_LH1_SB	SB	0	10	10	10	20	50	0.000	50	20–25, 25–30
18	2004_LH_SB	SB	10	20	30	30	30	50	0.000	60	0–15
19	2006_LH_SB	SB	10	30	30	40	80	50	0.244	45	10–12, 14–16
20	2008_Beehive_SB_1	SB	10	20	30	30	50	45	0.059	34	10–20, 60–80
21	2008_Beehive_SB_2	SB	0	20	20	40	80	50	0.147	34	60–80
22	2009_MtHutt_CT1	CT	2	8	12	14	19	21	0.000	30	0–10, 60–70
23	2009_MtHutt_CT_2	CT	10	13	13	14	17	21	0.000	25	40–70
24	2003_Campbell_Fig6	RB	2	3	4	5	6	6	0.222	63	15–20, 35–40
25	2004_Campbell_Fig9	RB	1	2	2	2	6	6	0.015	65	5–20, 30–35

*QLCT: quantified loaded column test (Landry and others, 2001); SF: shear frame (Perla and others, 1982); SB: stuffblock (Birkeland and Johnson, 1999); CT: compression test (Jamieson, 1999); RB: rutschblock (Föhn, 1987a).
†Defined threshold for a 'strong' stability test (shown in Table 1).
‡As shown in Figure 4.
§Probability that a measurement in the dataset is greater than our threshold of a 'strong' stability test.

chances of sampling two strong tests at our smallest sampling interval, 0–2 m.

The five datasets (17–21) using the stuffblock test are more difficult to compare because they use three different sampling layouts. Two interesting datasets are 20 and 21, both of which utilized a 10 m by 10 m sampling grid layout used by Hendrikx and others (2009). In both of these datasets, only a minimal chance of sampling two strong tests exists at any distance (Fig. 4; Table 3). This may be because spatial autocorrelation for these data only exists at distances less than our 10 m spacing.

The next two datasets (22 and 23) used the compression test and the same spatial layout as datasets 20 and 21 (used by Hendrikx and others, 2009). In these datasets the greatest distances from 40 to 70 m minimize the chances of sampling two strong tests; however, dataset 22 also has an additional minimum around 0–10 m (Fig. 4; Table 3).

The final two datasets (24 and 25) utilized the rutschblock test. Though the sampling layouts for these two datasets are not identical, spacing for the tests is similar. Longer distances (>30 m) helped to minimize the probability of two strong rutschblocks in both of these datasets, but an additional minimum for the first was evident at 15–20 m, while for the second that minimum existed at all distances from 5 to 20 m (Fig. 4; Table 3).

As an alternative to discussing the datasets by test type or layout, we can divide them by slope and snowpack characteristics (Table 2). Though complicated by variations in sampling layout and test type, this analysis is intended to see whether any distinct patterns related to the slope position or weak-layer properties emerge from our data. Unfortunately, we cannot find any clear and convincing pattern for our data. For example, the weak layer of interest in 15 of our datasets (60%) is surface hoar, and the distances that minimize the chances of sampling two strong tests in those datasets range from 0–2 m (dataset 12) up to 25–30 m (datasets 8, 9 and 17) (Tables 2 and 3). Faceted crystals comprised the weak layer in four (16%) of our datasets, and distances for minimizing the chances of two strong tests in these datasets ranged from 10 to 30 m. Likewise, the four (16%) datasets fracturing on depth hoar had distances ranging from 10 to nearly 80 m depending on the test type and sampling layout. We also had two datasets (8%) where the weak layer was decomposing fragments, and in these cases the distances ranged from around 10 to 70 m (Tables 2 and 3). Thus, no patterns exist related to weak-layer grain type. Dividing the datasets by whether they are above, at or below treeline presents an equally complicated picture, with a range of distances for each category. Likewise, binning the datasets by stability, measured as the probability

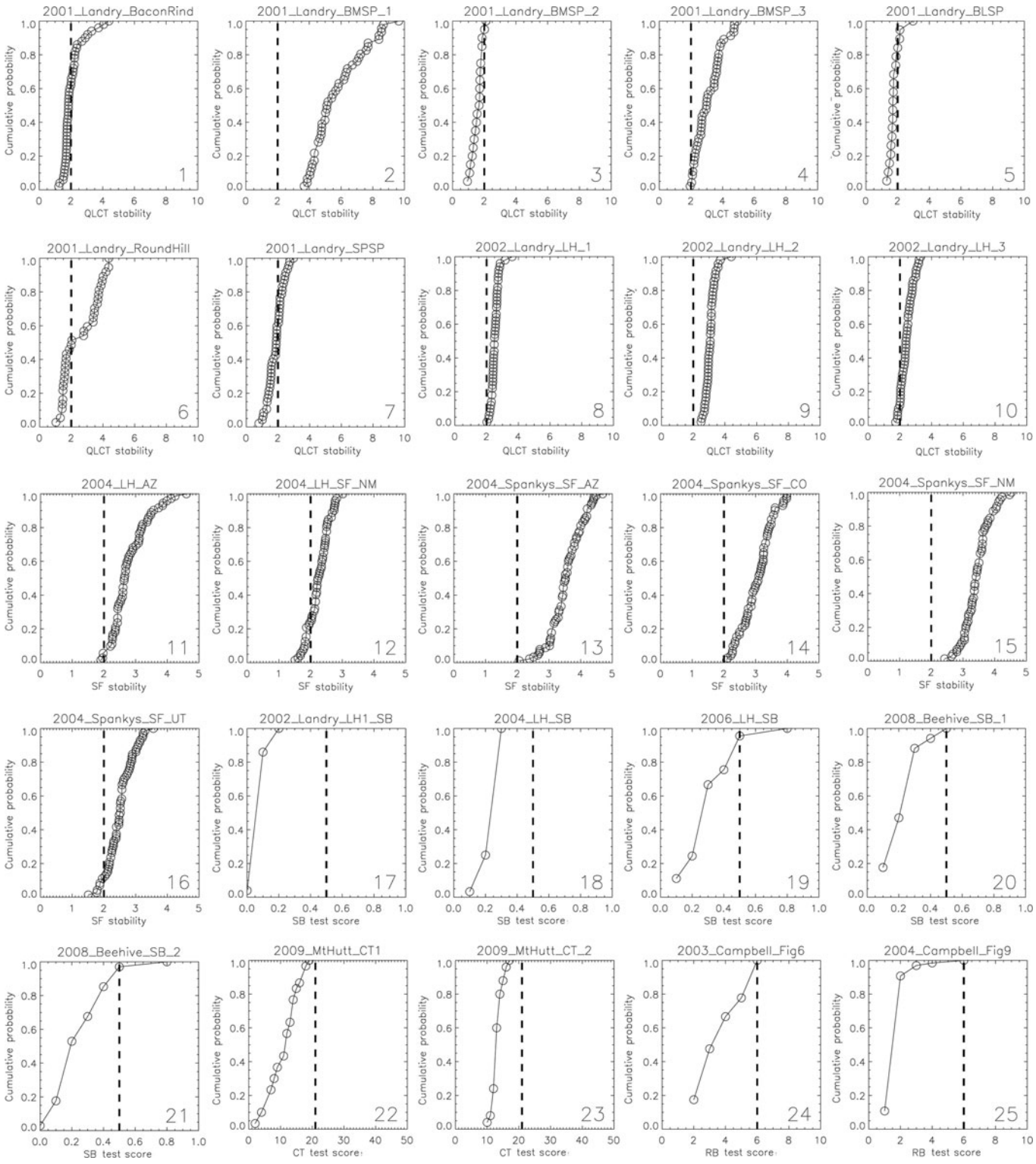

Fig. 2. CDF for each of our datasets. The prescribed stability thresholds are shown as vertical dashed lines.

that a test within that dataset is greater than or equal to our probability thresholds (Table 2), does not result in any easily identifiable patterns.

Most of our datasets (72%) come from slopes that are relatively unaffected by wind (Table 2). These datasets exhibit the entire range of distances that minimize the chances of two strong tests, from 0–2 m all the way up to 60–80 m (Table 3). Fewer datasets (28%) are from wind-affected sites. Six of these seven wind-affected datasets have distances greater than 10 m. This hints that it may be especially important to space stability tests at appropriately large distances on wind-affected slopes to avoid sampling two strong tests. However, this conclusion is based on a limited number of datasets using different sampling strategies and different tests, so it should be viewed with appropriate scientific skepticism.

Independent of the method for dividing our datasets, no clear patterns emerge and we cannot provide any concrete guidelines for test spacing. In most situations, it is better to space out tests by at least 5 m rather than put them close to

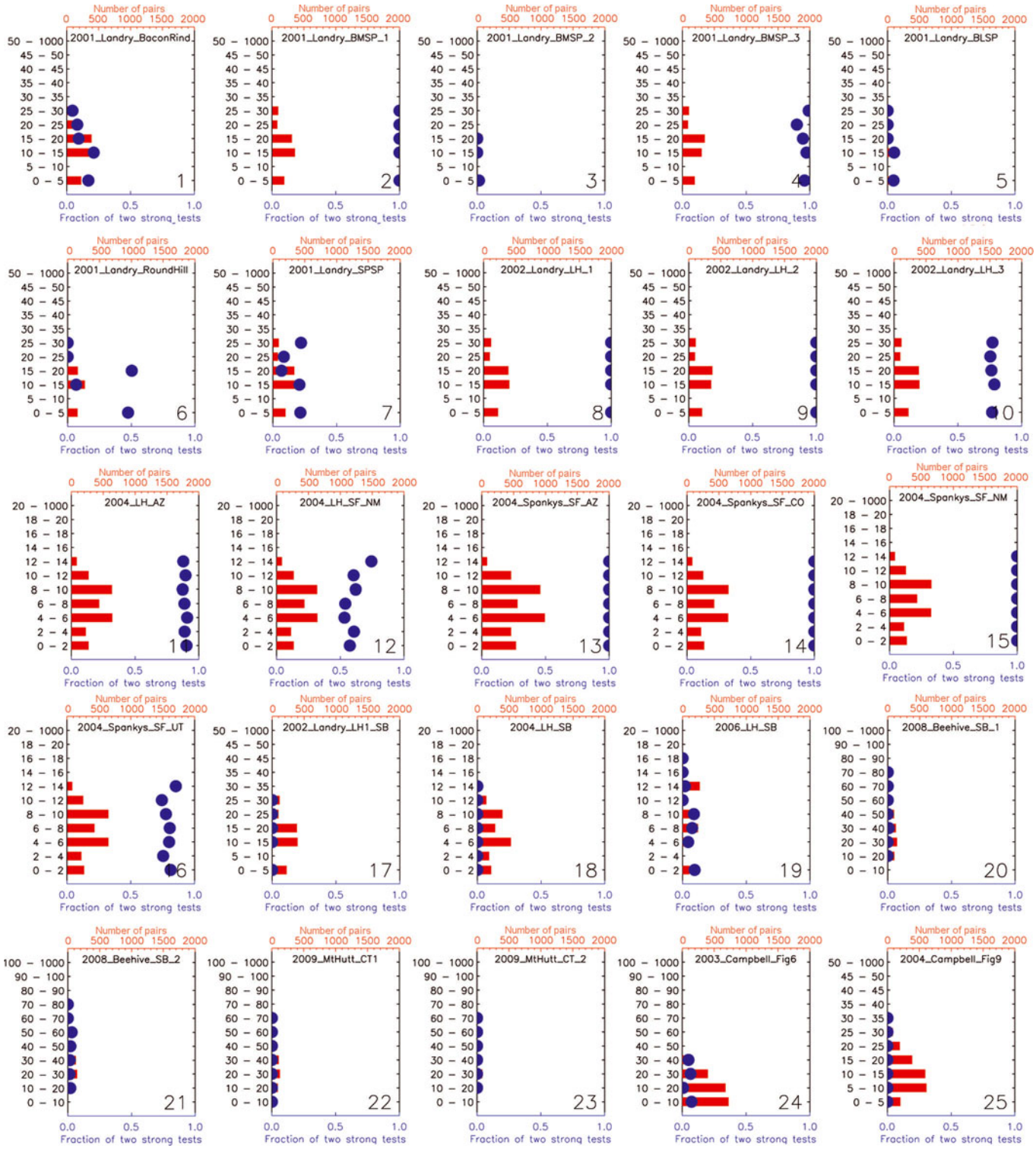

Fig. 3. The number of point pairs at each distance, and the fraction of two strong tests for each of our 25 datasets. A strong test result in this figure is defined as the thresholds shown in Table 3, and by the vertical dashed lines in Figure 2.

each other. For example, in 23 of our 25 datasets (88%) the optimal spacing of tests was >6 m. However, we did have three datasets (12%) where the optimal distance was <6 m, and the optimal distance to minimize the probability of two strong tests in our other datasets varied widely. In essence, our data suggest that the optimal distance will likely vary from slope to slope and from situation to situation.

The variability of our results is similar to the variations in autocorrelation lengths found in other spatial variability studies. For example, Kronholm and Schweizer (2003) and Kronholm and others (2004) quantified lengths varying from 2 to >10 m, Campbell (2004) and Campbell and Jamieson (2007) found lengths of 1–14 m, Birkeland and others (2004) showed lengths of 5–8 m, and Logan and others (2007) found little or no autocorrelation. An advantage of our work is that we do not look at a single autocorrelation length; rather, our analyses investigate the spatial range of the data to find the distance which minimizes the probability of sampling two strong tests.

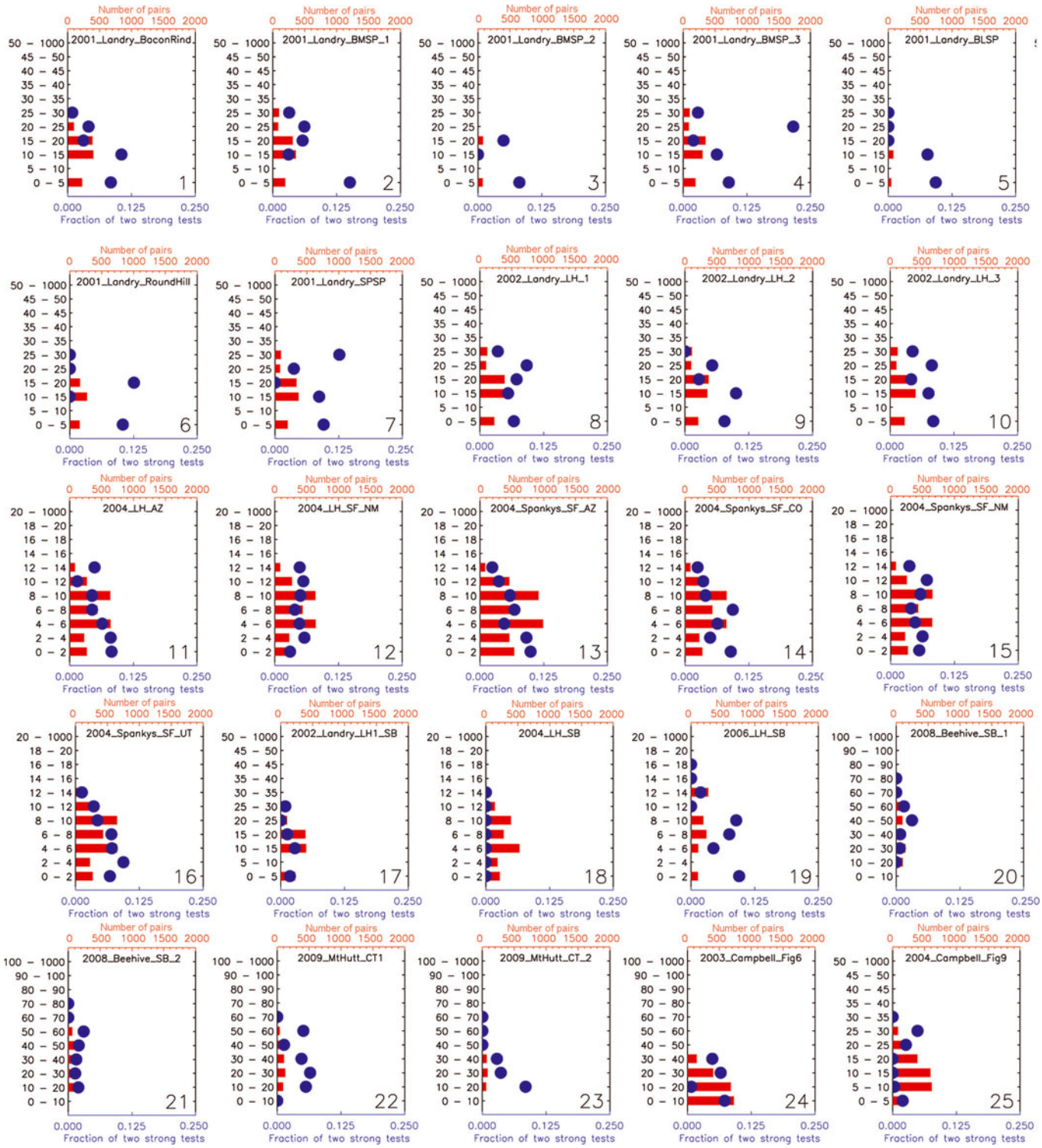

Fig. 4. The number of point pairs and the fraction of two strong tests for each of our 25 datasets. A strong test result in this figure is defined as being ≥75th percentile of the dataset, allowing us to more effectively explore the spatial variability of each dataset.

CONCLUSIONS

The optimal distance to space stability tests to minimize the probability of sampling two strong tests varies between our datasets and is independent of test type, spatial layout and weak-layer crystal type (Tables 2 and 3; Figs 3 and 4). Our results do show that this optimal distance is rarely <5 m; only two of our 25 datasets (8%) demonstrate this characteristic. This is mostly consistent with previous recommendations of spacing tests at least 10 m apart (Schweizer and others, 2008), and suggests that avalanche forecasters and other practitioners should not necessarily rely on two adjacent tests when searching for instability, but that a longer distance may help to reduce the probability of sampling two strong tests. However, the optimal distance is still an open question since we can see cases in our data where certain longer distances actually maximize our chances of measuring two strong tests. In fact, our work shows that there may be no such thing as an optimal

distance, but rather there are a range of suboptimal distances that one would like to avoid, and these vary from slope to slope and situation to situation.

It is not surprising that closely spaced tests generally do not minimize the probability of sampling two strong, or relatively strong, tests. Closely spaced tests should have similar aspect, slope angle, wind effect and snowpack structure and therefore would likely be similar. Of course, occasionally we also have fairly remarkable variation at these close distances; this is shown in some of our data (e.g. datasets 12, 18 and 22), as well as in some previous research (e.g. Landry and others, 2004; Campbell and Jamieson, 2007). However, even at the longer distances, we cannot provide guidance for spacing tests, since our results vary between datasets. This is also not unexpected. Each slope is unique and has different characteristics that are known to affect variability, such as slope substrate, wind patterns, snow depth, slight changes in aspect, and differences in energy balance across the slope that can affect weak layer formation and persistence (Birkeland and others, 1995; Campbell and Jamieson, 2007; Schweizer and others, 2008; Lutz, 2009).

Improved procedures for spatial analyses might provide more conclusive results. Unfortunately, this is a difficult task when utilizing classic snow stability tests. There is a limit to the number of data collectable in a day, especially if one observer conducts the tests to minimize observer variability. Further, collecting data over a period longer than 1 day is likely to introduce temporal changes into the spatial analysis due to the rapidly changing snowpack. Perhaps other measurement techniques (e.g. radar) will provide larger datasets, but currently such data only quantify snow structure and not snow stability (Marshall and Koh, 2008).

From a practical perspective, the variability that exists on slopes increases the uncertainty associated with our stability assessments. One way to help lower this uncertainty is to collect multiple stability tests from different parts of the slope in a search for instability. Though it is generally better to space these tests some distance, the optimal spacing will vary from slope to slope as well as from situation to situation. Thus, experienced observers are critically important for the collection and interpretation of good data. Of course, ultimately a holistic approach is required, whereby the experienced observer takes into account not only stability-test results, but also weather, avalanche and snowpack observations to assess the avalanche potential.

ACKNOWLEDGEMENTS

Numerous individuals helped collect field data, including C. Landry, K. Kronholm, E. Lutz, S. Logan, R. Johnson, J. Chipman, P. Staron, J. Nelson and T. Chesley. The Gallatin National Forest Avalanche Center provided logistical support for our Montana data collection. We thank C. Campbell and B. Jamieson for the use of the Canadian rutschblock data. Partial funding for this work came through a Fulbright Senior Specialist Grant, the Royal Society of New Zealand International Science and Technology (ISAT) Linkage Fund, the US National Science Foundation (grant BCS-024310), the New Zealand Foundation for Research Science and Technology (C01X0812), and New Zealand National Institute of Water and Atmospheric Research (NIWA) Capability funding. We thank two anonymous reviewers for providing useful comments that improved the paper.

REFERENCES

Birkeland, K.W. and D. Chabot. 2006. Minimizing 'false-stable' stability test results: why digging more snowpits is a good idea. *In* Gleason, J.A., *ed. Proceedings of the International Snow Science Workshop, 1–6 October 2006, Telluride, Colorado*. Telluride, CO, International Snow Science Workshop, 498–504.

Birkeland, K.W. and R.F. Johnson. 1999. The stuffblock snow stability test: comparability with the rutschblock, usefulness in different snow climates, and repeatability between observers. *Cold Reg. Sci. Technol.,* **30**(1), 115–123.

Birkeland, K.W., K.J. Hansen and R.L. Brown. 1995. The spatial variability of snow resistance on potential avalanche slopes. *J. Glaciol.,* **41**(137), 183–190.

Birkeland, K., K. Kronholm and S. Logan. 2004. A comparison of the spatial structure of the penetration resistance of snow layers in two different snow climates. *In* Ganju, A., *ed. Proceedings of the International Symposium on Snow Monitoring and Avalanches, 12–16 April 2004, Manali, India*. Manali, Snow and Avalanche Study Establishment, 3–11.

Blöschl, G. 1999. Scaling issues in snow hydrology. *Hydrol. Process.,* **13**(14–15), 2149–2175.

Campbell, C.P. 2004. Spatial variability of slab stability and fracture properties in avalanche start zones. (MS thesis, University of Calgary.)

Campbell, C. and B. Jamieson. 2007. Spatial variability of slab stability and fracture characteristics within avalanche start zones. *Cold Reg. Sci. Technol.,* **47**(1–2), 134–147.

Föhn, P.M.B. 1987a. The 'Rutschblock' as a practical tool for slope stability evaluation. *IAHS Publ.* 162 (Symposium at Davos 1986 – *Avalanche Formation, Movement and Effects*), 223–228.

Föhn, P.M.B. 1987b. The stability index and various triggering mechanisms. *IAHS Publ.* 162 (Symposium at Davos 1986 – *Avalanche Formation, Movement and Effects*), 195–214.

Fredston, J.A. and D. Fesler. 1994. *Snow sense: a guide to evaluating snow avalanche hazard.* Anchorage, AK, Alaska Mountain Safety Center.

Gauthier, D. and B. Jamieson. 2008. Fracture propagation propensity in relation to snow slab avalanche release: validating the propagation saw test. *Geophys. Res. Lett.,* **35**(13), L13501. (10.1029/2008GL034245.)

Greene, E. *and 10 others*. 2009. *Snow, weather, and avalanches: observational guidelines for avalanche programs in the United States. Second edition*. Pagosa Springs, CO, American Avalanche Association.

Hendrikx, J. and K.W. Birkeland. 2009. Spatial variability and the extended column test: results from Mount Hutt. *Crystal Ball,* **18**(3), 17–20.

Hendrikx, J., K. Birkeland and M. Clark. 2009. Assessing changes in the spatial variability of the snowpack fracture propagation propensity over time. *Cold Reg. Sci. Technol.,* **56**(2–3), 152–160.

Jamieson, J.B. 1999. The compression test – after 25 years. *Avalanche Rev.,* **18**(1), 10–12.

Kronholm, K. and K.W. Birkeland. 2007. Reliability of sampling designs for spatial snow surveys. *Comput. Geosci.,* **33**(9), 1097–1110.

Kronholm, K. and J. Schweizer. 2003. Snow stability variation on small slopes. *Cold Reg. Sci. Technol.,* **37**(3), 453–465.

Kronholm, K., M. Schneebeli and J. Schweizer. 2004. Spatial variability of micropenetration resistance in snow layers on a small slope. *Ann. Glaciol.,* **38**, 202–208.

Landry, C.C., J. Borkowski and R.L. Brown. 2001. Quantified loaded column stability test: mechanics, procedure, sample-size selection, and trials. *Cold Reg. Sci. Technol.,* **33**(2–3), 103–121.

Landry, C.C., K. Birkeland, K. Hansen, J. Borkowski, R.L. Brown and R. Aspinall. 2004. Variations in snow strength and stability on uniform slopes. *Cold Reg. Sci. Technol.,* **39**(2–3), 205–218.

Logan, S., K. Birkeland, K. Kronholm and K. Hansen. 2007. Temporal changes in the slope–scale spatial variability of the shear strength of buried surface hoar layers. *Cold Reg. Sci. Technol.,* **47**(1–2), 148–158.

Lutz, E.R. 2009. Spatial and temporal analysis of snowpack strength and stability and environmental determinants on an inclined forest opening. (PhD thesis, Montana State University.)

Marshall, H.-P. and G. Koh. 2008. FMCW radars for snow research. *Cold Reg. Sci. Technol.,* **52**(2), 118–131.

McClung, D. and P. Schaerer. 2006. *The avalanche handbook. Third edition.* Seattle, WA, The Mountaineers.

Perla, R., T.M.H. Beck and T.T. Cheng. 1982. The shear strength index of alpine snow. *Cold Reg. Sci. Technol.,* **6**(1), 11–20.

Schweizer, J. and S. Bellaire. 2009. Where to dig? On optimizing sampling strategy. *In* Schweizer, J. and C. Gansner, *eds. Proceedings of the International Snow Science Workshop, 27 September–2 October 2009, Davos, Switzerland.* Birmensdorf, Swiss Federal Institute for Forest, Snow and Landscape Research, 298–300.

Schweizer, J., K. Kronholm, J.B. Jamieson and K.W. Birkeland. 2008. Review of spatial variability of snowpack properties and its importance for avalanche formation. *Cold Reg. Sci. Technol.,* **51**(2–3), 253–272.

Simenhois, R. and K.W. Birkeland. 2009. The extended column test: test effectiveness, spatial variability, and comparison with the propagation saw test. *Cold Reg. Sci. Technol.,* **59**(2–3), 210–216.

Tremper, B. 2008. *Staying alive in avalanche terrain. Second edition.* Seattle, WA, The Mountaineers Books.

Voight, B. and 15 others. 1990. Snow avalanche hazards and mitigation in the United States. Washington, DC, National Academy Press.

Webster, R. and M.A. Oliver. 2001. *Geostatistics for environmental scientists.* Chichester, Wiley.

MS received 1 February 2010 and accepted in revised form 19 July 2010