Cold Regions Science and Technology 64 (2010) 110–118

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

Statistical evaluation of local to regional snowpack stability using simulated snow-cover data

Michael Schirmer *, Jürg Schweizer, Michael Lehning

WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, 7260 Davos, Switzerland

ARTICLE INFO

Article history:
Received 13 November 2009
Accepted 22 April 2010

Keywords:
Avalanche forecasting
Snow stability
Avalanche danger
Snow cover
Numerical modelling

ABSTRACT

Snow stability, or the probability of avalanche release, is one of the key factors defining avalanche danger. Most snow stability evaluations are based on field observations, which are time-consuming and sometimes dangerous. Through numerical modelling of the snow cover stratigraphy, the problem of having sparsely measured regional stability information can be overcome. In this study we compared numerical model output with observed stability. Overall, 775 snow profiles combined with Rutschblock scores and release types for the area surrounding five weather stations were rated into three stability classes. Snow stratigraphy data were then produced for the locations of these five weather stations using the snow cover model SNOWPACK. We observed that (i) an existing physically based stability interpretation implemented in SNOWPACK was applicable for regional stability evaluation; (ii) modelled variables equivalent to those manually observed variables found to be significantly discriminatory with regard to stability, did not demonstrated equal strength of classification; (iii) additional modelled variables that cannot be measured in the field discriminated well between stability categories. Finally, with objective feature selection, a set of variables was chosen to establish an optimal link between the modelled snow stratigraphy data and the stability rating through the use of classification trees. Cross-validation was then used to assess the quality of the classification trees. A true skill statistic of 0.5 and 0.4 was achieved by two models that detected "rather stable" or "rather unstable" conditions, respectively. The interpretation derived could be further developed into a support tool for avalanche warning services for the prediction of regional avalanche danger.

1. Introduction

The European avalanche danger scale is defined based upon snowpack stability (i.e. the probability of avalanche release), the frequency of trigger points (spatial distribution of instability) and the size and type of the anticipated avalanches (Meister, 1995). Stability is the only property that can be estimated through observations and interpretations of snow profiles and stability tests (Schweizer and Wiesinger, 2001; Schweizer et al., 2008). Since observations are time-consuming and sometimes dangerous, avalanche warning services sparsely receive information about snowpack stability. Numerical modelling of snow cover stratigraphy and stability has been proposed as a solution to this problem.

The question arises how reliable are evaluations provided by numerical models. Durand et al. (1999) compared their modelled stability estimate of the SAFRAN/Crocus/MÉPRA (SCM) chain to observed avalanche activity. Despite the fact that two different datasets could not be strictly compared, the forecast quality was similar to what is typically achieved with weather information alone (Heierli et al., 2004; Pozdnoukhov et al., 2008). The link between avalanche activity and corresponding danger level has previously been shown to be inconsistent. Schweizer et al. (2003) reported this discrepancy was mainly due to limited visibility during periods of high activity. Furthermore, said study concluded that avalanche occurrence data are not suitable to verify lower danger levels ("Low", "Moderate" and "Considerable"). Schirmer et al. (2009) linked statistically simulated snow cover data to forecasted avalanche danger. It was shown that simulated snow cover information was useful for statistical danger level prediction and provided additional benefit in comparison to weather information alone. Lehning et al. (2004) summarised the stability evaluations implemented in the snow cover model SNOWPACK and assessed their quality in comparison to the forecasted avalanche danger level. Schweizer et al. (2006) developed a new stability evaluation based on SNOWPACK simulations obtaining critical thresholds between three stability classes. However, those results were not cross-validated because of a limited dataset of $N=33$. Hence the question of how well this evaluation might perform on an independent dataset could not be answered.

The first aim of this study was to assess the quality of existing stability estimates implemented in SNOWPACK using a large dataset of observed stability. The second aim was to test modelled variables

* Corresponding author. Fax: +41 81 4170 110.
E-mail address: schirmer@slf.ch (M. Schirmer).

0165-232X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.coldregions.2010.04.012

equivalent to manually observed variables (Schweizer and Jamieson, 2003) found to be significantly discriminating between stable and unstable observed profiles. The third aim was to detect additional modelled variables which discriminate well between stability categories. These variables were then used to link modelled stratigraphy to observed stability through the use of classification trees.

In addition to other automatic methods potentially supporting avalanche danger forecasting (for detecting avalanche days with weather data (e.g. Buser, 1983; Heierli et al., 2004; Pozdnoukhov et al., 2008), for predicting the avalanche danger level itself using measured (Schweizer and Föhn, 1996) or simulated snow stratigraphy (Schirmer et al., 2009)), the approach developed in this study for estimating stability, covers a supplementary facet of the avalanche danger prediction process, which could be used as a support tool for avalanche warning services.

2. Methods

2.1. Overview

The main purpose of this study was to relate modelled snow cover data to measured stability information. A summary of the methodology follows:

1. 775 stability observations from one region were used as target variable, which were interpreted with two (one subjective and one objective) rating schemes.
2. Modelled snow cover information evaluated at five automatic weather stations in that region was used to explain the observed stability.
3. Due to the large amount of modelled snow cover information a reduction of data was applied:
   (a) Only modelled properties of the slab, the weak layer and the surface were considered together with measured and modelled meteorological variables.
   (b) The large amount of remaining variables (~300) were rated with the Fisher criterion to select the 20 best uncorrelated variables.
4. Classification trees were trained with the 20 best variables as input to explain observed stability. Similarly, univariate trees were built in order to test stability estimates already implemented in the snow cover model.
5. The results achieved from the model were validated through comparison to expert knowledge and agreement with observations.
6. Additionally, a discussion on whether the classification trees could be used for a probability forecast is presented.

2.2. Data

In order to relate stability observations with simulated snow cover data using automatic weather stations (AWS) as input, a test region was chosen, where many snow profiles with stability tests in the surrounding of weather stations were available. An analysis of the SLF snow profile database showed that only the region of Davos, in the Eastern Swiss Alps, had a sufficient number of stability observations for a statistical analysis. Five AWS are located in the region. We selected observations within a 5 km radius and an elevation band of $\pm$300 m of the stations. These thresholds were chosen to consider the two following conflicting aspects: (i) increasing the dataset would make the statistical analysis more reliable, while (ii) observations at larger distances to the AWS might have less relation to the simulated snow cover data. We obtained 775 cases where both the stability observation and the simulated snow cover were available.

The observations included a Rutschblock test and a snow profile. Since the Rutschblock score is dependent on the inclination, we considered only observations from slopes of an inclination >20°. Observations with snow depth less than 50 cm were not included, as it was assumed that with this restriction the stability interpretation would be more reliable. Since we were mainly interested in dry snow situations, only observations between November and April were considered.

The observations were rated into the three stability classes “poor”, “fair” and “good”. This was achieved through the application of two existing stability interpretations. The first of which is a subjective interpretation scheme developed with expert knowledge (Schweizer and Wiesinger, 2001). Expert knowledge is also needed to apply this rating. For the dataset used in this study profiles were rated by different people. We must therefore assume that the rating is not fully consistent, on the other hand, the expert is able to include a broad spectrum of information in the rating. The second rating is an objective, rule-based method. This method was statistically developed in trying to find differences between observations recorded on slopes that were adjacent to skier-triggered avalanches (“unstable”) and those that were skied but not triggered (“stable”) (Schweizer et al., 2008). Applying both interpretations, a relatively similar distribution of the three categories was obtained (30% “poor”, 30% “fair” and 40% “good”). However, in only 60% of the cases did the ratings agree. Slight differences in the distribution indicated that the objective interpretation is more conservative (tends to produce more unstable ratings).

The snow cover model SNOWPACK was used to generate the corresponding snow stratigraphy (Bartelt and Lehning, 2002; Lehning et al., 2002a,b). The model provides a huge amount of data in high temporal resolution. Therefore we reduced the data by considering mainly failure layer and slab properties. The stability index (*SSI*) developed by Schweizer et al. (2006) defined the potential weak layer interface in the modelled snow cover. Similar to a study of observed stratigraphy, the softer layer was chosen as failure layer and the harder as adjacent layer (Schweizer and Jamieson, 2003). These modelled stratigraphy variables were completed with meteorological and snow-surface variables (e.g. measured wind velocity or modelled surface albedo). In addition to modelled values at noon, we considered also sum, mean, extreme values or rate, for different time intervals. However, this leads to an increase in the number of possible variables making a reduction of the variables even more necessary.

We evaluated existent stability evaluations implemented in SNOWPACK, which are mainly stability indices relating parameterised shear strength with shear stress. Those are Sk_{38} (Jamieson and Johnston, 1998) and *SSI* delivering continuous values, while a combination of both, which is abbreviated as “SC” (Schweizer et al., 2006), delivers the three stability categories “poor”, “fair” and “good”.

2.3. Rating of variables

A simple univariate rating was performed using the Fisher criterion, which is defined as the ratio of the between-class variance to the within-class variance (e.g. Bishop, 2006) and for a two class problem is given by

$$J = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \quad (1)$$

where m_i is the mean and s_i the standard deviation of class i, $i = 1,2$. As input for the classification trees described below the 20 best non-pairwise linearly correlated variables were considered ($r^2 < 0.6$).

2.4. Classification

Classification trees (Breiman et al., 1998) were used to discriminate between the stability categories. The number of stability categories were reduced to minimise the dimensionality of the forecast verification problem. Murphy (1991) defined the dimensionality of a forecast with the number of quantities needed to reconstruct the joint distribution of forecast and observation. Since the absolute forecast verification problem

of a three-category forecast is already eight dimensional (square of number of categories − 1), the (interrelated, but) multi-facetted nature of forecast quality will be easier to obtain when the number of categories is reduced (Murphy, 1991; Murphy, 1993). Therefore we trained and verified the trees not on the three categories of the stability observation ("poor", "fair" and "good"). Instead, trees were built for the detection of "poor" observations, and other trees for the detection of "good" observations, while the remaining two categories were pooled.

The classification trees were obtained by optimising the misclassification costs and the complexity (size) of a tree (Breiman et al., 1998). Furthermore, altered prior probabilities were used to adjust the individual class misclassification. Choosing a prior probability larger than the observed relative frequency will tend to decrease the misclassification of that class. This will also introduce a bias between modelled and observed class relations in form of a higher modelled relative frequency (Breiman et al., 1998). We decided to adjust priors to achieve a forecast with larger values for another — for this study more important — quality characteristic (skill, definition in Section 2.5 and in Table 3), which outweighed the disadvantage of the bias in class relations. Through varying the prior probabilities we optimised the true skill statistic (Doswell et al., 1990). Due to the previously mentioned difference in characteristics between modelled and observed classes, from here on "good" and "poor" conditions will be referred to as "rather stable" and "rather unstable", respectively. Both model parameters, tree size and priors, were determined through cross-validation.

2.5. Validation

Two distinct types of goodness for a forecast system were considered as identified by Murphy (1993) (Table 1): The first is the correspondence between forecasts and judgements. This is the so-called type 1 goodness or consistency. We evaluated whether the objective data analysis led to a model which was mechanistically consistent with existing (physical) process understanding as expressed by the current expert knowledge. A discussion will be presented later as to whether the tree branching can be reasonably explained and whether the rating of the input variables with the Fisher criterion is comparable to the experts' rating. Consequently, we asked five experts to take a survey to select up to ten different modelled variables which they believed to discriminate between stable and unstable conditions.

The second type of goodness is the straightforward agreement between the forecasts and corresponding observations (i.e. the type 2 goodness or quality). This was assessed through cross-validation (CV). The quality of a forecast can be assessed with the joint distribution of forecast f and observation x, which can be displayed in terms of a 2×2 contingency table in the case of a binary categorical forecast (Table 2). Although the table is comprehensive, the information is more accessible when factorised in conditional distributions (Murphy and Winkler, 1987). In order to determine how well the forecast discriminates between observation classes, the probability of detection (POD) and the probability of false detection (POFD) were chosen (Table 3). The frequency of correct null forecasts (FOCN) and the false alarm ratio (FAR) deliver additional insight into how reliable forecasts are. These measures are sample estimates of the conditional distributions. To inspect further aspects of quality, we chose to highlight the accuracy expressed with the proportion correct (PC) and the skill

Table 1
Types of goodness of a forecast (Murphy, 1993).

Name	Description
Consistency	Type 1 goodness, consistency with existing (physical) process understanding.
Quality	Type 2 goodness, agreement between forecasts and corresponding observations, which can be expressed with measures listed in Table 3.

Table 2
Contingency table for a binary forecast ("1": event, "0": non-event). Total number of cases: $N=a+b+c+d$.

		Observation x	
		1	0
Forecast f	1	a	b
	0	c	d

of a forecast (Wilks, 1995). The skill of a forecast is defined as the relative accuracy with respect to a standard reference forecast. This reference forecast is random and unbiased for the true skill statistic (TSS), which is obtained by subtracting POFD from POD (Wilks, 1995). An overview of the evaluation quantities used is given in Table 3.

2.6. Probability forecast

Classification trees can be applied to create a probabilistic forecast using the class relations at a terminal node. Each terminal node i is now denoted as a separate forecast f_i. The corresponding forecast probability $p(f_i)$ can be obtained from the class relations by applying a large dataset. Since prior probabilities were used, these class relations need to be adjusted to calculate the forecast probability for each node:

$$p(f_i,j) = \frac{\pi_j N_{ij} / N_j}{\sum_k \pi_k N_{ik} / N_k}, \tag{2}$$

where $p(f_i, j)$ is the forecast probability for the forecast f_i and class j, π_j the prior probability, N_j the initial class frequency and N_{ij} the class frequency at node i for class j, while $k=1,2$.

The quality of a probabilistic forecast can be assessed with an attribute diagram (Wilks, 1995). An attribute diagram relates the forecast probability $p(f_i, j=1)$, obtained from the training parts of the CV blocks, to the observed relative frequency $p(x=1|f_i)$ obtained from the test parts of the CV blocks. These quantities were calculated for the same trees and with the same cross-validation blocks, which were used for the verification of the categorical forecast (classification, see Section 2.4). The attribute diagram delivers insight into quality aspects as reliability and resolution. Reliability of a probability forecast is the correspondence between forecast probability and the observed relative frequency, while resolution is the ability of the forecasts to sort the observed events into groups that are different from each other (Wilks, 1995). Results in terms of reliability and resolution are discussed in Section 3.2.

2.7. Suitability of proposed methods

One single stability observation has only limited strength of explanation for a regional evaluation. Schweizer et al. (2003) presented

Table 3
Quality measures (Doswell et al., 1990; Wilks, 1995).

Measure	Description
POD (probability of detection)	Probability that event "1" was forecasted when it was observed, $p(f=1\|x=1)$. Estimated with $a/(a+c)$.
POFD (probability of false detection)	Probability that event "1" was forecasted when it was not observed, $p(f=1\|x=0)$. Estimated with $b/(b+d)$.
FOCN (frequency of correct null forecast)	Probability that the non-event "0" was observed when it was forecasted, $p(x=0\|f=0)$. Estimated with $d/(c+d)$.
FAR (false alarm ratio)	Probability that the event "1" was not observed when it was forecasted, $p(x=0\|f=1)$. Estimated with $b/(a+b)$.
PC (proportion correct)	Probability that the observed event was correctly forecasted. Estimated with $(a+d)/N$.
TSS (true skill statistic)	Measure of skill. Skill is the relative accuracy with respect to a reference forecast. Estimated with POD-POFD $=(ad-bc)/(a+c)(b+d)$.

characteristic stability distributions for different danger levels. They concluded that only a sufficient number of observations allows a reliable regional stability estimate. However, the dataset in this study would not have contained enough examples if only days that had such a sufficient number of observations were considered. Therefore, days with only one stability observation were also included. The question then was whether a learning system is able to adapt rules correctly, when the stability observations, in some cases, are of limited value for the true prediction parameter "regional stability"? We assumed that there are some limited combinations of factors explaining regional stability and that these patterns repeat in time. An example being snow fall in combination with large wind speeds, or a weak layer of surface hoar crystals together with certain slab properties. Both of these may be such repeating patterns, which have an influence on regional stability. Consequently, even though this assumption was only partly fulfilled, it made sense to include days to the dataset with only one stability observation.

In the dataset used for this study, which contains 775 stability observations, 314 observations were unique per station and day. On average, 1.8 stability observations were available per day and per station. Multiple observations associated with one station on a single day were sometimes inconsistent due to the in-region variability or the uncertainty of the observation. Since simulated snow cover data were only available once per station, multiple observations were applied in the learning phase, although sometimes inconsistent, to the same modelled input variables of each day. This was intended to filter out the in-region effects and help find the main factors of repeating patterns influencing the regional stability. In the evaluation phase, median values of observations made on the same day were used, which where weighted with the number of observations per day, since the classification trees were not able to reproduce the in-region variance of observed stability. The rare cases with median values between two classes were neglected, which lead to the varying number of cases in Tables 5 and 6 compared to the original dataset ($N=775$).

A model validation is more reliable if a better estimate of the regional stability can be used. Over several periods of time a verified regional avalanche danger level, based on a sufficient number of observations, was available for the region of this study (Schweizer et al., 2003; Schweizer and Kronholm, 2007). Therefore, the performance of the trees for these reduced time periods were independently tested.

Modelled snow cover variables are auto-correlated over long periods of time (Schirmer et al., 2009), which prohibited random cross-validation (Elsner and Schmertmann, 1994). Modelled values at nearby days (both past and future) introduce a bias to the estimation of forecast skill, as these values contain noise correlated to noise for the omitted day. Furthermore, nearby future days were likely to be especially informative about omitted target variables. However, they would be unavailable in real forecast situations, which would bias the estimate of the forecast skill towards higher values. Therefore, blocks of data that were uncorrelated in time were removed. In the case of this study, it was necessary to select blocks for entire winters. We applied such a winter-by-winter CV to select input variables, model parameters (tree size and prior probabilities) and to evaluate quality aspects. One has to keep in mind that the cross-validated quality of a classification tree through CV is assessed by dissimilar trees, which are built with the training parts of the cross-validations blocks. These trees differ in input variables, nodes and tree size, both with respect to each other and to the classification tree for which the quality is validated.

3. Results

3.1. Rating of variables

The rating performed with the Fisher criterion (Eq. 1) was applied twice, first for the detection of the category "poor" and second for "good". For the detection of the category "good" higher values of the Fisher criterion were achieved. Higher values were also obtained when the subjective stability interpretation was used for classifying the manual observations in comparison to the objective interpretation. Furthermore, quality characteristics of classification trees trained on stability observations rated with the subjective interpretation were better in comparison to the objective interpretation. Subsequently, we will only show results obtained from the subjective stability interpretation.

The two stability indices Sk_{38} and *SSI* implemented in SNOWPACK showed different strength of discrimination. For both the detection of rather stable or rather unstable conditions, the values of the Fisher criterion for Sk_{38} were larger than for *SSI*. Similar results were achieved with a Kruskal–Wallis test for the three stability categories: The Sk_{38} was able to discriminate between the three categories ($p<0.001$), while the *SSI* showed no significant strength of discrimination ($p>0.05$). The combination of the indices ("SC") was also not significant.

For some of the variables which discriminated well between stable and unstable profiles for observed stratigraphy, i.e. failure layer grain size, hardness, and their differences to the adjacent layer (Schweizer and Jamieson, 2003) the sign of correlation was different for modelled profiles. For example, while in unstable observed profiles the difference in grain size across the failure interface was typically large, it was small in modelled profiles. This indicates a low consistency between model and observations.

For other significant variables in the observed stratigraphy, the signs of correlation agreed. Rather unstable profiles had shallower snow depth and lower failure layer shear strength than rather stable profiles. They were typically classified as profile types 7 and 4 (while rather stable profiles were classified as profile type 6) (Schweizer and Lütschg, 2001). Depth hoar was more often found in failure layers of rather unstable profiles (in modelled stratigraphy also faceted crystals), while rounded grains were more often found in stable profiles.

In contrast to the observations found by Schweizer and Jamieson (2003), the slab was significantly thicker in rather unstable modelled profiles. Slab density, one of the most important modelled variables according to the Fisher criterion (lower densities corresponded to more unstable profiles), was not significant in observed stratigraphy. However, there are some indications that soft slabs are more dangerous (Schweizer and Lütschg, 2001; Schweizer and Jamieson, 2003).

Variables representing a change in time were rarely rated as important. This suggests that manual snow profiles contain the most relevant information or that snowpack changes are slow.

As input for classification trees, only the 20 best not pair wise linearly correlated variables were considered ($r^2<0.6$) and are listed in Table 4. The majority were selected both for the detection of rather stable and rather unstable conditions. Most selected variables are modelled. Exceptions are wind speed and surface temperature, which were measured at the weather station.

As mentioned in Section 2.7, experts were independently asked to select modelled variables which should discriminate between rather stable and rather unstable conditions. If selected variables were also chosen by experts, they are marked in Table 4. For most other variables selected by experts high values of the Fisher criterion were obtained. Exceptions were, for example, the difference in hardness or density between failure and adjacent layer, an increase in air temperature in the last 24 h or the existence of a crust in the slab, which were chosen by experts but achieved low values of the Fisher criterion.

3.2. Classification

The trees using as input the best variables shown in Table 4 are presented in Fig. 1a for the detection of the rather stable conditions and in Fig. 1b, for the detection of rather unstable conditions (model "Best_20"). To evaluate the first type of goodness of a forecast as

Table 4
The 20 best pair wise uncorrelated variables selected with the Fisher criterion for the detection of rather stable conditions and rather unstable conditions. The abbreviation "diffminmax" stands for absolute difference between maximum and minimum value of the mentioned time interval.

Variable	Selected for rather stable	Selected for rather unstable	Selected by experts
Mean slab density	x	x	x
Mean slab bond size	x	x	x
Difference in hand hardness between slab and failure layer	x	x	x
Ratio of mean slab bond size to grain size	x	x	x
Failure layer bond size times grain size	x	x	x
Ratio of failure layer bond size to grain size	x	x	x
Failure layer 3D coordination number (Lehning et al., 2002a)	x	x	x
Depth hoar within 1 m beneath the penetration depth	x	x	x
3 day new snow sum	x	x	x
3 day new snow sum (24 hour diffminmax)	x	x	
24 hour new snow sum (24 hour diffminmax)	x	x	
24 hour maximum of 3 hour new snow sum	x	x	
Ski penetration depth	x	x	
24 hour mean wind speed	x	x	x
Sensible heat flux (24 hour diffminmax)	x	x	
Surface temperature (24 hour diffminmax)	x	x	
Snow temperature at 10 cm below the surface	x	x	
12 hour rate of snow temperature at 10 cm below the surface	x		
Strain rate of adjacent layer (Lehning et al., 2002a)	x		x
Failure layer contains surface hoar	x		x
Ski penetration depth (24 hour diffminmax)		x	
24 hour mean of energy fluxes at surface		x	
Adjacent layer consists of faceted crystals		x	

introduced above, the consistency, we discuss the physical or logical meaning of the presented classification trees.

For the detection of rather stable conditions (Fig. 1a) the most important variable was the 3-day new snow sum. On the one hand, it seems logical for many situations that no new snow is related to rather stable conditions as suggested by the tree. On the other hand, in the subjective interpretation scheme (which was applied here to rate the manual observations) no reference to the new snow amount is given. However, the new snow amount will affect other measured quantities such as hardness or the Rutschblock score, which are used by this stability interpretation. The next node used the ratio of failure layer bond size to grain size, where large values are related to rather stable conditions as large bond size in comparison to grain size suggests large shear strength. At the next node the tree classified the conditions as rather stable if depth hoar was not present within a depth of 1 m below ski penetration. This is appropriate since failures in depth hoar layers deeper than 1 m below the ski penetration is unlikely (Schweizer and Camponovo, 2001). Weak layers up to this depth may potentially be released by skiers (cp. Schweizer and Jamieson, 2003) and depth hoar is related to low shear strength (Jamieson and Johnston, 2001).

For the detection of rather unstable conditions (Fig. 1b) the most important variable selected by the tree was the ski penetration depth. Larger values were attributed to rather unstable conditions, which is consistent with the observations mentioned previously that soft slabs are related to unstable conditions. In many situations large penetration depth corresponded to new snow conditions, which points to rather unstable conditions. Additionally, a large penetration depth facilitates triggering of deep weak layers. The next nodes (on the right hand side of Fig. 1b) have previously been discussed. On the left hand side the next node is defined by the absolute difference between the minimum and the maximum of the snow surface temperature T_{ss} in the last 24 h. Low values are related to cloudy conditions, which can explain the relation to unstable conditions. The tree suggests from the last node on the left hand side rather unstable conditions if the failure layer bond size is small. Small bond size should be related to low values of shear strength, hence rather unstable conditions.

In interpreting of the sub nodes in the tree hierarchy, an important consideration is that the detected dependencies are, in fact, only valid for the subgroup of the dataset reaching this node: the conditions of the preceding nodes have to be applied first. However, all dependencies detected for these subgroups were validated through testing against the whole dataset. In conclusion, the nodes and threshold values are plausible and the trees displayed a high degree of consistency when compared to the judgement of the experts.

Tables 5 and 6 present the cross-validated contingency tables for the model "Best_20". In both cases, i.e. the detection of rather stable and unstable conditions, event "1" ("good" for the detection of rather stable, "poor" for the detection of rather unstable) was forecasted more frequently than it was observed (bias larger than one (Wilks, 1995)). This was artificially produced with the altered prior probabilities, as was mentioned in Section 2.4. Recall, the prior probabilities were optimised so as not to obtain an unbiased forecast, rather a forecast with the largest skill; with the consequence that the event "1" is more frequently predicted than observed.

The corresponding performance measures are visualised in Fig. 2, together with the models "Sk_{38}" and "SC". Results for trees using only the input variable *SSI* are not shown, since their performance measures were not as good as trees using the Sk_{38}. In the detection of rather stable conditions (Fig. 2a), a proportion correct (PC) of 0.75 was obtained for the model "Best_20". This high value has only limited implication, as a constant forecast of "poor/fair" would achieve a PC of ~0.7 since the dataset is unbalanced with a base rate of 0.34. Such a constant forecast would have no skill, whereas the model "Best_20" achieved a true skill statistic (TSS) of 0.5. This model was able to discriminate between the observations, with a probability of detection (POD) of 0.76 and a probability of false detection (POFD) of just 0.26. The model was also quite reliable as when "poor/fair" was forecasted, it was correct with a probability of 0.86 (FOCN, frequency of correct null forecasts). The relatively high value (0.4) of the false alarm ratio (FAR) indicates some deficits in reliability. In all aspects of quality discussed, the model "Best_20" performed better than the models "Sk_{38}" and "SC". However, the differences to the "Sk_{38}" were small. The "Sk_{38}" model achieved a lower PC (0.70), a lower POD (0.72), a lower TSS (0.42), a lower FOCN (0.83) and higher FAR (0.44).

The quality of the detection of rather unstable conditions can be assessed with Fig. 2b. While the POD for the "Best_20" model was again quite large (0.77), the relatively large value for POFD (0.37) resulted in a lower TSS (0.4) than for the detection of rather stable conditions. While the model was very reliable when "fair/good" was forecasted (FOCN = 0.9), the same level of reliability was not present when "poor" was forecasted (FAR = 0.6). Again, for all aspects of quality discussed, the model "Best_20" performed better than the other two models.

In summary, the detection of the rather stable conditions was easier to detect, which is consistent with larger values of the Fisher criterion. Reasons for the low performance measures for the "SC" model can be found in the small dataset used in Schweizer et al. (2006) ($N = 33$). The published non cross-validated accuracy could not be achieved when cross-validation was applied using our large dataset.

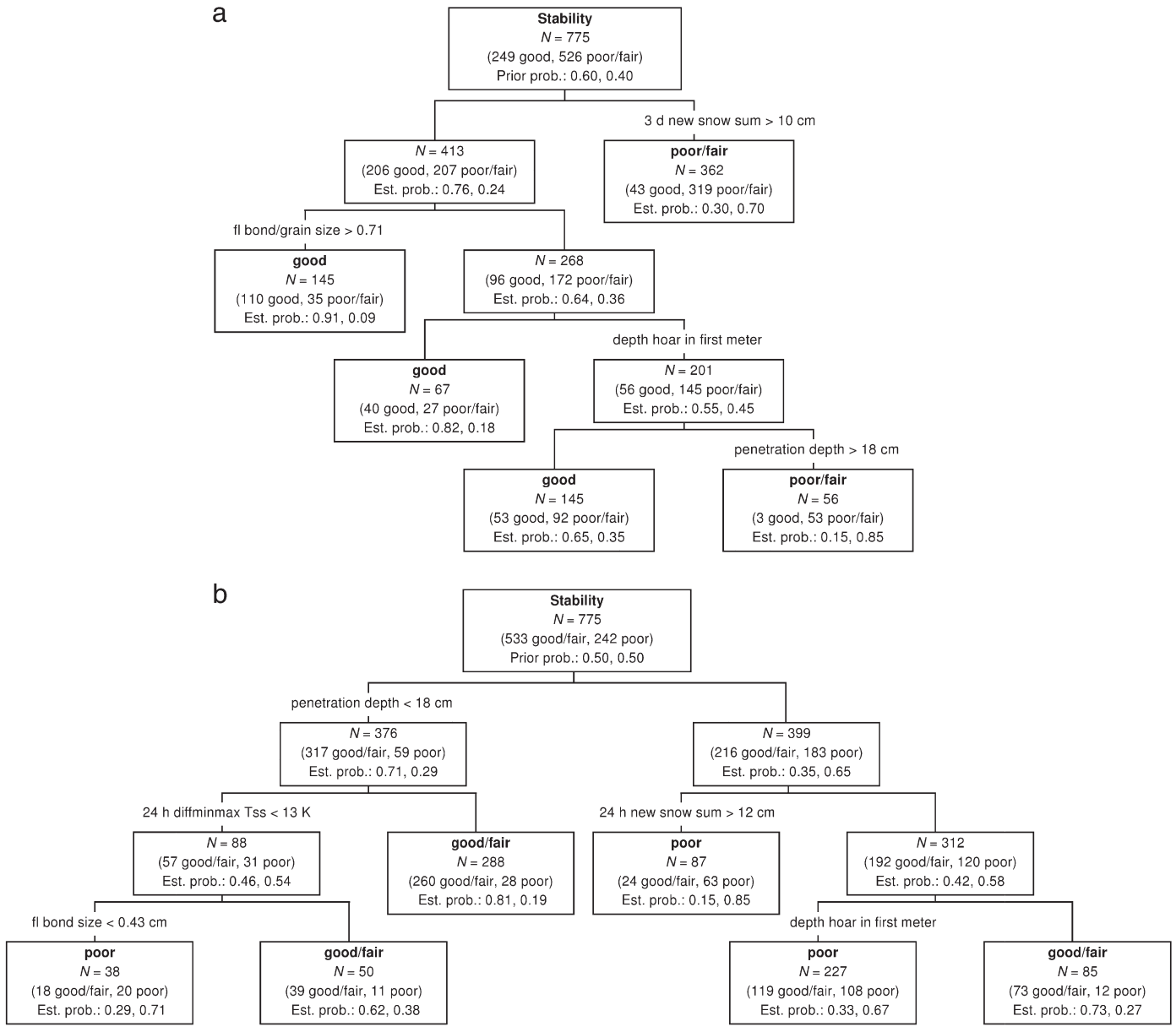

Fig. 1. Classification tree for the detection of (a) rather stable and (b) unstable conditions using the 20 best uncorrelated variables defined with the Fisher criterion ("Best_20"). For each node the count *N* of examples reaching that node, the class relations and the estimated forecast probabilities ("Est. prob.") are recorded. In each first node, the values of the altered prior probabilities ("Prior prob.") are noted. Failure layer is abbreviated with "fl", temperature of the snow surface with "Tss" and the absolute difference between minimum and maximum with "diffminmax".

The univariate tree for Sk_{38} had a single split, which suggests rather unstable conditions for values smaller than 0.58 and rather stable conditions for larger values.

As discussed in Section 2.7, a model validation would benefit from a verified regional stability estimate. There are several days for which a reliable verified avalanche danger level could be used for validation, which is shown in Table 7 for the model "Best_20". This was done for the categorical (non-probabilistic) forecast. However, what can be expected when modelled stability estimates are compared to avalanche danger levels? Referring to the typical stability distributions found by Schweizer et al. (2003), the tree which detects the rather stable conditions must detect the danger level "1: Low": the

Table 5
Contingency table of the model "Best_20" for the detection of rather stable conditions. The base rate (fraction of observations of class "1") was 0.34.

		Observation x		
		1: good	0: poor/fair	Total
Forecast f	1: good	190	128	318
	0: poor/fair	61	364	425
	Total	251	492	743

Table 6
Contingency table of the model "Best_20" for the detection of rather unstable conditions. The base rate (fraction of observations of class "1") was 0.25.

		Observation x		
		1: poor	0: fair/good	Total
Forecast f	1: poor	139	201	340
	0: fair/good	41	336	377
	Total	180	537	717

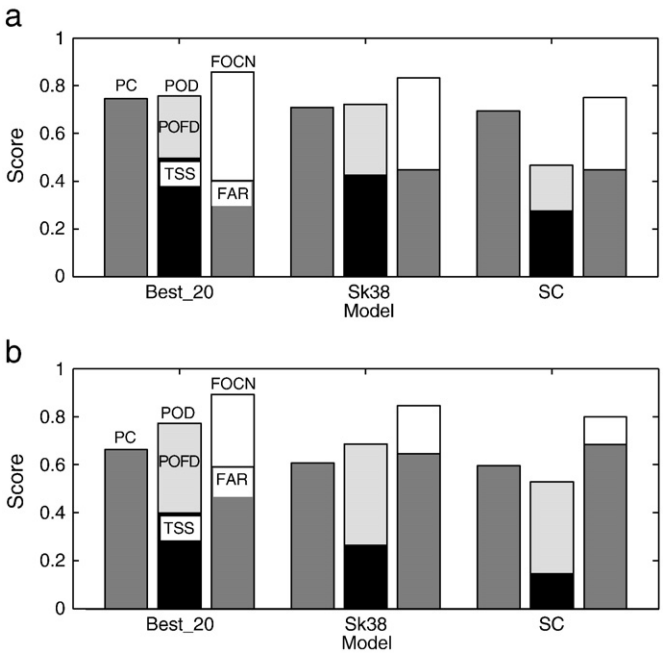

Fig. 2. Overview of the models' quality measures for the detection of (a) rather stable and (b) rather unstable conditions. Presented are the proportion correct (PC), the probability of detection (POD), the probability of false detection (POFD) and the true skill statistic (TSS). Furthermore, the frequency of correct null forecasts (FOCN) and the false alarm ratio (FAR) are also shown.

binary distribution "good" vs. "poor/fair" is 90% to 10% for level "Low" and 25% to 75% for "2: Moderate". The situation for the tree which detects rather unstable conditions is not as clear: at danger level "3: Considerable", the typical distribution of the binary variable "poor" vs. "fair/good" is mostly balanced. However, since this tree produced more "poor" situations than are present in the observations (bias of 1.5), it is expected that the detection of rather unstable conditions might correspond to the detection of danger levels greater than or equal to "Considerable".

Considering these arguments, on 11–12 December 2002 the tree to detect rather unstable conditions failed. The verified danger level was "Low", but the tree output was "poor", which corresponds to danger levels greater than or equal to "Considerable". On the other hand, the tree to detect rather stable conditions delivered the output "good", which is consistent with the verified danger level. The inconsistent output of the two trees ("poor" vs. "good") can be interpreted as an indication of uncertainty. In the next period the verified danger level was "Moderate", hence the tree to detect rather unstable conditions should not show "poor", since the avalanche danger level is lower than "Considerable". This was correctly reproduced. Similarly, the tree for the detection of rather stable conditions should not show "good", since the danger level was not "Low", which is also correctly reproduced. All cases for which the models did not correspond to the verified danger level are marked with an asterisk in Table 7 (three cases in twenty). There were no instances where both models are wrong. By comparison, for five in ten cases the forecasted danger level was one level different from the verified danger level. However, it is necessary to recall that the classification trees were used as a nowcast, while the forecasted danger level was issued on the afternoon of the preceding day.

Table 7
Results of the cross-validated classification trees using the best 20 variables as input ("Best_20") applied to periods with verified regional danger level (Schweizer et al., 2003; Schweizer and Kronholm, 2007). Cases for which the models did not correspond to the verified danger level were marked with an asterisk.

Date	Forecasted danger level	Verified danger level	Tree to detect "poor"	Tree to detect "good"
11–12 Dec 2002	2	1	Poor*	Good
21–23 Jan 2002	1	2	Fair/good	Poor/fair
12–13 Feb 2002	3	3	Poor	Poor/fair
26–27 Feb 2002	3	3	Poor	Poor/fair
18–19 Mar 2002	2	1–2	Fair/good	Poor/fair*
20 Mar 2002	3	3	Poor	Poor/fair
7 Jan 2003	2	3	Poor	Good*
13 and 15–17 Jan 2003	2	1	Fair/good	Good
7 Feb 2003	4	3–4	Poor	Poor/fair
17–20 Feb 2003	2	1	Fair/good	Good

3.3. Probability forecast

It was tested whether the classification trees of the "Best_20" could be used for a probabilistic forecast. Fig. 3 shows the attribute diagram (Wilks, 1995), which relates the forecast probability to the observed relative frequency. A circle for each separate forecast determined with the terminal nodes of the trees for each cross-validation block is drawn. Some terminal nodes were more frequently used when the test dataset was applied. This frequency is expressed with the size of the circles. The reference to the categorical use of the trees (classification) is given as follows: the examples of the test dataset with a forecast probability larger than the prior probability would have been classified as event "1". In the detection of rather stable conditions (Fig. 1a) a dependency between forecast probability and observed relative frequency is recognisable. Frequently used nodes with larger forecast probability do show relatively more observations of the category "good". This is visualised with larger circles close to the perfect reliability line (iii). It can be seen that examples classified as "poor/fair" also more often contributed positively to the forecast skill; they are more often in the grey zone, while many examples, which would be classified as "good", are in the white zone thus contributing negatively to the forecast skill. This is consistent with the verification of the categorical forecast showing large FOCN values and intermediate FAR values. Only one very frequently used node showed perfect reliability.

Consistent with the observations for the categorical forecast, a lower quality was observed for the detection of rather unstable conditions (Fig. 3b); many circles lie outside the grey zone, hence contributing negatively to the forecast skill. Again, examples with low forecast probabilities contributed positively to the skill more frequently.

4. Discussion and conclusion

The first objective of this study was to determine if existing physically based stability estimates implemented in SNOWPACK were applicable for regional stability assessment. Based on the Fisher criterion, this is not the case for the stability index *SSI*, while for the Sk_{38} this was confirmed. Such a result seems plausible since the *SSI* uses the variables difference in hardness and grain size to adjust the Sk_{38}. These variables were implemented under the hypothesis that modelled variables equivalent to those identified as significant for observed stratigraphy would show similar strength of classification. However, this hypothesis now seems incorrect, since for the most important variables in observed stratigraphy an opposite sign of correlation to stability was found for modelled variables. Those variables were difference in hardness and grain size across the failure interface, failure layer grain size and hardness. Nevertheless, for some variables an agreement between modelled and observed snow stratigraphy variables and their relation to stability was found, e.g. grain type and weak layer shear strength. Of particular importance is that other variables, which cannot be measured or were not considered so far (e.g. mean properties of the slab and differences of these to weak layer properties), did show a good strength of classification.

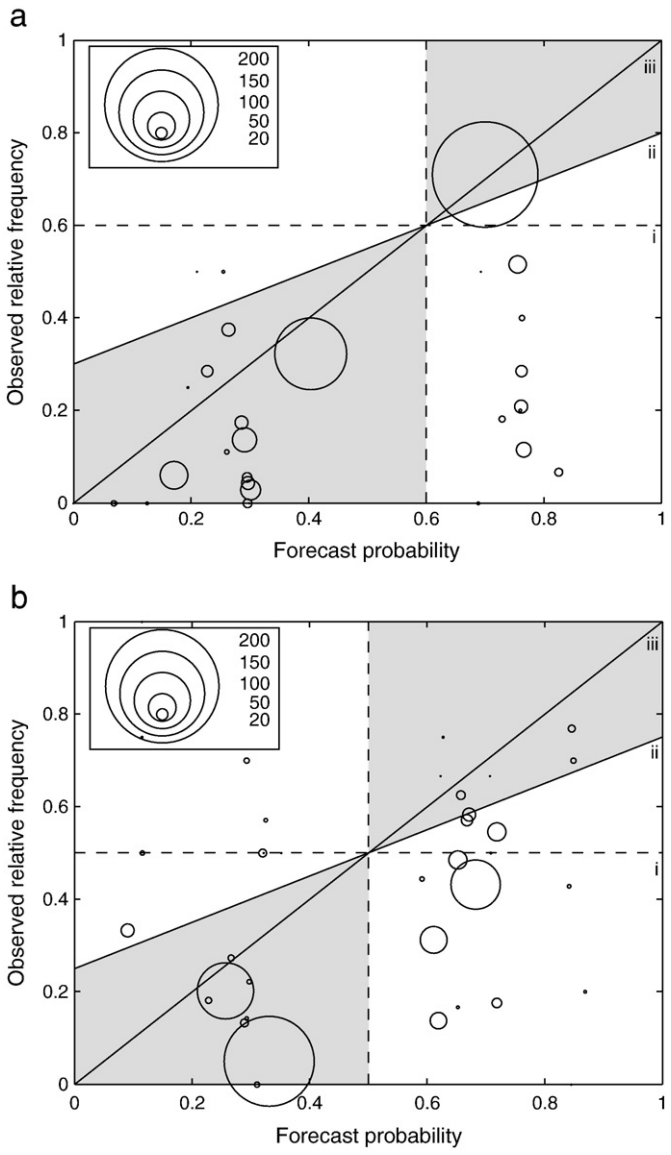

Fig. 3. Attribute diagram for the verification of the probabilistic forecast of model "Best_20" for the detection of (a) rather stable and (b) rather unstable conditions. It relates the forecast probability of each node of the trees obtained from the training parts of the CV blocks and the observed relative frequency obtained from the test parts. The size of the circles shows how often the nodes were used by the test parts. Circles on line (i) have no resolution, which is plotted at the level of the prior probability for class "1". Circles on line (ii) indicates no skill, while line (iii) implies perfect reliability and skill. Circles in the grey zone contribute positively to forecast skill.

The majority of important variables were from the modelled snow stratigraphy, rather than the measured meteorological data (Table 4). This suggests that snowpack modelling delivers an additional benefit to the measured input variables when snow stability needs to be estimated, which is consistent with the findings of Schirmer et al. (2009).

As the *SSI* was used for failure layer detection, it was based upon modelled differences in hardness and grain size. In contrast, the *SSI* was not able to distinguish between stable and unstable conditions, though it was constructed to both find a relevant weak layer and give a stability estimate (Schweizer et al., 2006). In fact, convincing failure layers and interfaces were detected, e.g. buried surface hoar, depth hoar and faceted crystals as well as crusts. Furthermore, modelled properties of these layers were important to distinguish between stable and unstable conditions, e.g. the Sk_{38} but not the *SSI* itself. Thus it seems that detecting a failure layer and estimating stability are different processes requiring different variables. In the case of the observed stratigraphy, this is confirmed by a study performed by Schweizer and Jamieson (2007), which recognised that a method developed for stability estimation can only be partially adapted for detecting potential failure layers.

Several forecast qualities, i.e. discrimination, reliability and skill (TSS = 0.5 and 0.4) were computed for the classification trees presented in Section 3.2, as overall performance cannot be expressed by a single quality measure. The classification trees using the most important variables as selected by the Fisher criterion performed better than stability estimates already implemented in SNOWPACK. When periods with a verified regional danger level were tested the trees performed convincingly as only three in 20 cases the models failed. The probabilistic forecast provided by the classification trees can be used with limitations, at least for the detection of rather stable conditions a reliable forecast quality was observed.

Since separate classification trees were used to detect rather stable and rather unstable conditions, potentially conflicting tree results can occur (Table 7). Classification trees applied on the original three categories achieved lower quality characteristics, which was to be expected with the higher dimensionality of the classification problem. Furthermore, the two class problem simplified the validation task (as discussed in Section 2.4). It is believed that the disadvantage of sometimes conflicting tree results is outweighed by the advantages of (i) a higher performance of the separate trees and (ii) the simplified validation. Additionally, conflicting results can be used as an indication of uncertainty in addition to the estimated probability delivered by the classification trees.

Consistency with expert knowledge was found: the variables independently selected for the model by experts were often rated as important with the Fisher criterion. Physical and logical explanations of the rules delivered by the classification trees were found.

Our models for stability evaluation could be used by avalanche warning services as nowcast or forecast, though only in regions with similar climatic characteristics as in the region of Davos. It is characterised by a transitional snow climate, with a maximum snow depth of 2 m snow in average, 2500 m above sea level. The region of Davos exhibits climatological differences at the five stations used. Since the classification trees were trained without location specific information for where the snow cover variables were generated, the climatological differences were integrated in the learning process; only those rules could be detected, which were valid for all five stations. Thus, it is believed that the methods presented can be applied to many regions without further modifications. However, for regions with other characteristics, a new classification tree must be trained. This is only possible if a similarly large amount of observations are available as was the case for this study.

Nowcasting may already have an important value, since information on instability at the present day is rare. In a forecasting operation, the present snow cover would be simulated with measured data first and then the development of the snow cover would be predicted with forecasted meteorological data for the next day. Finally, the predicted snow cover would provide the additional input variables needed for the classification trees. The uncertainty of the forecasted input parameters and its effect on the classification trees was not assessed in this study.

Acknowledgements

This work has partially been funded by the Swiss National Science Foundation and the Swiss Federal Office of the Environment. We acknowledge the valuable comments provided by Jake Turner and two anonymous reviewers. This study would not have been possible without our colleagues at SLF who made field observations in the last ten years, especially Roland Meister and the avalanche warning service. We are very grateful to Charles Fierz, Sascha Bellaire and

Christoph Mitterer for their insightful commentaries on the ongoing work and together with Walter Steinkogler, Christine Groot Zwaaftink and Vanessa Wirz for their contribution to the expert survey.

References

Bartelt, P., Lehning, M., 2002. A physical SNOWPACK model for the Swiss avalanche warning. Part I: numerical model. Cold Regions Science and Technology 35 (3), 123–145.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1998. Classification and Regression Trees. CRC Press, Boca Raton, U.S.A.

Buser, O., 1983. Avalanche forecast with the method of nearest neighbours: an interactive approach. Cold Regions Science and Technology 8 (2), 155–163.

Doswell, A., Davies-Jones, R., Keller, D., 1990. On summary measures of skill in rare event forecasting based on contingency tables. Weather and Forecasting 5, 576–585.

Durand, Y., Giraud, G., Brun, E., Mérindol, L., Martin, E., 1999. A computer-based system simulating snowpack structures as a tool for regional avalanche forecasting. Journal of Glaciology 45 (151), 469–484.

Elsner, J., Schmertmann, C., 1994. Assessing forecast skill through cross validation. Weather and Forecasting 9, 619–624.

Heierli, J., Purves, R., Felber, A., Kowalski, J., 2004. Verification of nearest neighbours interpretations in avalanche forecasting. Annals of Glaciology 38 (5), 84–88.

Jamieson, B., Johnston, C.D., 2001. Evaluation of the shear frame test for weak snowpack layers. Annals of Glaciology 32 (11), 59–69.

Jamieson, J., Johnston, C., 1998. Refinements to the stability index for skier-triggered dry-slab avalanches. Annals of Glaciology 26, 296–302.

Lehning, M., Bartelt, P., Brown, B., Fierz, C., Satyawali, P., 2002a. A physical SNOWPACK model for the Swiss avalanche warning. Part II: snow microstructure. Cold Regions Science and Technology 35 (3), 147–167.

Lehning, M., Bartelt, P., Brown, B., Fierz, C., 2002b. A physical SNOWPACK model for the Swiss avalanche warning. Part III: meteorological forcing, thin layer formation and evaluation. Cold Regions Science and Technology 35 (3), 169–184.

Lehning, M., Fierz, C., Brown, B., Jamieson, B., 2004. Modeling snow instability with the snow-cover model SNOWPACK. Annals of Glaciology 38 (8), 331–338.

Meister, R., 1995. Country-wide avalanche warning in Switzerland. Proceedings of the International Snow Science Workshop (ISSW 1994), 30 October–3 November 1994. Snowbird, Utah, pp. 58–71.

Murphy, A., 1991. Forecast verification: its complexity and dimensionality. Monthly Weather Review 119 (7), 1590–1601.

Murphy, A., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8, 281–293.

Murphy, A., Winkler, R., 1987. A general framework for forecast verification. Monthly Weather Review 115 (7), 1330–1338.

Pozdnoukhov, A., Purves, R., Kanevski, M., 2008. Applying machine learning methods to avalanche forecasting. Annals of Glaciology 49 (7), 107–113.

Schirmer, M., Lehning, M., Schweizer, J., 2009. Statistical forecasting of regional avalanche danger using simulated snow-cover data. Journal of Glaciology 55 (193), 761–768.

Schweizer, J., Bellaire, S., Fierz, C., Lehning, M., Pielmeier, C., 2006. Evaluating and improving the stability predictions of the snow cover model SNOWPACK. Cold Regions Science and Technology 46 (1), 52–59.

Schweizer, J., Camponovo, C., 2001. The skier's zone of influence in triggering slab avalanches. Annals of Glaciology 32 (7), 314–320.

Schweizer, J., Föhn, P., 1996. Avalanche forecasting — an expert system approach. Journal of Glaciology 42 (141), 318–332.

Schweizer, J., Jamieson, B., 2003. Snowpack properties for snow profile interpretation. Cold Regions and Science Technology 37 (3), 233–241.

Schweizer, J., Jamieson, J., 2007. A threshold sum approach to stability evaluation fo manual snow profiles. Cold Regions Science and Technology 47 (1–2), 50–59.

Schweizer, J., Kronholm, K., 2007. Snow cover spatial variability at multiple scales: characteristics of a layer of buried surface hoar. Cold Regions Science and Technology 47 (3), 207–223.

Schweizer, J., Kronholm, K., Wiesinger, T., 2003. Verification of regional snowpack stability and avalanche danger. Cold Regions Science and Technology 37 (3), 277–288.

Schweizer, J., Lütschg, M., 2001. Characteristics of human-triggered avalanches. Cold Regions Science and Technology 33 (2–3), 147–162.

Schweizer, J., McCammon, I., Jamieson, J., 2008. Snowpack observations and fracture concepts for skier-triggering of dry-snow slab avalanches. Cold Regions Science and Technology 51 (2–3), 112–121.

Schweizer, J., Wiesinger, T., 2001. Snow profile interpretation for stability evaluation. Cold Regions Science and Technology 33 (2–3), 179–188.

Wilks, D., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press, San Diego.