Regional stability evaluation with modelled snow cover data

Michael Schirmer *, Jürg Schweizer, Michael Lehning
WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland

ABSTRACT: Stability (or the probability of avalanche occurrence or release) is probably the key factor defining the avalanche danger level. Most snow stability evaluations are based on field measurements which are time-consuming and sometimes dangerous. Numerical modelling of snow cover stratigraphy and stability offers a solution to the problem of having only sparse information about the regional stability. We compared numerical model output with observed stability. Over 700 snow profiles combined with rutschblock score and release type in the surroundings of five weather stations were rated into three stability classes. Snow stratigraphy data were produced for the locations of these five weather stations using the snow cover model SNOWPACK. We determined whether (i) existing physically based stability interpretations implemented in SNOWPACK are applicable for regional stability evaluation; (ii) modelled variables equivalent to those, which were identified as significant for real snow covers, contain similar classification power; (iii) additional modelled variables, which cannot be measured in the field perform better. Finally, with objective feature selection a set of variables was chosen to obtain an optimal link between the modelled snow stratigraphy data and the stability rating using classification trees. Cross-validation was used to assess the quality of the classification trees. Another aspect of goodness, the consistency with experts' judgement, was considered. The interpretation derived can be further developed into a support tool for avalanche warning services to predict the regional avalanche danger.

KEYWORDS: Avalanche forecasting, snow stability, avalanche danger, snow cover, numerical modelling

1 INTRODUCTION

The European avalanche danger scale is defined based on the factors snowpack stability (i.e. the probability of avalanche release), the frequency of trigger points (or spatial distribution of instability) and the size and type of the anticipated avalanches (Meister, 1995). Stability is the only quantity, which can be estimated from measurements using stability interpretations which combine snow profiles and stability tests (Schweizer and Wiesinger, 2001; Schweizer et al., 2008). Because measurements are time-consuming and sometimes dangerous, avalanche warning services receive only sparse information about snowpack stability. Numerical modelling of snow cover stratigraphy and stability would be a solution to this problem.

But 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. However, Schweizer et al. (2003) reported that avalanche observations were not consistent with danger ratings, mainly due to limited visibility during periods of high activity. They concluded that avalanche occurrence data would not be suitable to verify lower danger levels (1-3). Lehning et al. (2004) summarized the stability evaluations implemented in SNOWPACK and related their quality to the forecasted avalanche danger level. Schweizer et al. (2006) developed a new stability evaluation based on SNOWPACK simulations and obtained critical thresholds between three stability classes. However, results were not cross-validated because of the limited dataset of $N = 33$. The question how good this evaluation might be on an independent dataset could not be answered.

The purpose of this study was to assess the quality of existing stability estimates implemented in SNOWPACK using a large dataset of observed stability. It was furthermore tested, if observed snow stratigraphy parameters found to be relevant for stability evaluation (Schweizer and Jamieson, 2003) are also important for the simulated snow cover.

Linking statistically simulated snow cover data to forecasted avalanche danger, Schirmer et al. (in press) showed that simulated snow cover information is useful for statistical danger level prediction. In our approach, this information was used to explain measured stability observations with classification trees. In addition to other automatic methods, for example statistical

Corresponding author address:
Michael Schirmer, WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos Dorf, Switzerland;
tel: +41 81 417 0282; fax: +41 81 417 0110;
email: schirmer@slf.ch

methods for detecting avalanche days (e.g. Buser, 1983), predicting the avalanche danger level itself using measurements (e.g. Schweizer and Föhn, 1996) or simulated snow cover data (Schirmer et al., in press), our new approach covers with a stability evaluation a supplementary facet of the avalanche danger prediction process, and can be used as a support tool for avalanche warning services.

2 METHODS

2.1 Data

In order to relate measured stability observations with simulated snow cover data using automatic weather stations (AWS) as input, a test region was chosen, where many measurements in the surrounding of weather stations were available. An analysis of SLF's snow profile database showed that only in the region of Davos in the Eastern Swiss Alps enough measurements for a statistical analysis were available. Five AWS are located in the region. We selected measurements in a distance of up to 5 km and within an elevation band of ±300 m to the stations. These thresholds were chosen to optimise the two following aspects: (i) Increasing the dataset would make the statistical analysis more reliable, while (ii) measurements at larger distances to the AWS might be less related to the simulated snow cover data. We obtained over 700 cases, where both the measured stability observation and the simulated snow cover were available.

The measurements comprised a Rutschblock test and a snow profile. Since the Rutschblock score is dependent on the inclination, we considered only measurements from slopes >20°. No measurements with snow depth lower than 50 cm were selected, because we assumed that with this restriction the stability interpretation of the measurements would be more reliable. Because we were mainly interested in dry snow situations, only measurements between November and April were considered.

These measurements were rated into three stability classes ('poor', 'fair' and 'good') applying two existing stability interpretations. The first is a subjective interpretation scheme developed with expert knowledge (Schweizer and Wiesinger, 2001). The second rating is an objective, rule-based method, which was statistically developed trying to find differences in the observations performed on slopes that were adjacent to skier-triggered avalanches ('unstable'), or that were skied but not triggered ('stable') (Schweizer et al., 2008).

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 these 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 the study of real snowpacks, the softer layer was chosen as failure layer and the harder as adjacent layer (Schweizer and Jamieson, 2003). These model variables were completed with measured and calculated meteorological and snow-surface variables (e.g. wind velocity or surface albedo).

For many variables it may make sense to also consider – besides their midday values – their sum, mean, extreme values or rate, for different time intervals. This leads to a rapid increase in the number of possible variables, which makes a reduction of the variables necessary.

The existent stability evaluations implemented in SNOWPACK, which we wanted to verify, are mainly stability indices relating shear strength with shear stress. We focused on the Sk_{38} and the *SSI* and a combination of both (Schweizer et al., 2006).

2.2 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

$$J = \frac{\left(m_2 - m_1\right)^2}{{s_1}^2 + {s_2}^2} \qquad (1)$$

where m_i is the mean and s_i the standard deviation of class i, i = 1,2.

With this rating it was determined which of the derived variables (mean etc.) should be used for further analysis. For each basic SNOWPACK parameter one derived variable representing a status (e.g. a mean) and one which describes a change in time (e.g. a rate) were selected.

This objective rating was complemented with a survey of five experts – both in snow-modelling and field work. The experts were asked to select up to ten different modelled variables which should discriminate between more stable and more unstable conditions.

2.3 Classification and evaluation

Classification trees were used to discriminate between the stability categories (Breiman

et al., 1998). The correspondence between the forecasts and matching observations was assessed through cross-validation (CV). The autocorrelation of modelled snow cover data prohibited random CV (Elsner and Schmertmann, 1994). Therefore blocks of data were removed, which were de-correlated in time, in our case blocks of a whole winter needed to be selected.

At some days and in some regions several measured stability observations were available. Since simulated snow cover data were only available once per region (and since we did not study time-dependent variations at one day), the developed classification trees were not able to reproduce the in-region variance of observed stability. Therefore we also applied a CV using median target values of each day. The rare cases with median values between two classes were neglected.

To simplify the classification and the verification problem (the latter is already eight dimensional with three categories), we trained and verified the trees not on the three categories of the target variable ('poor', 'fair' and 'good'). Instead, trees were built for the detection of 'poor' observations (rather unstable conditions), and other trees for the detection of 'good' observations (rather stable conditions). We assessed the forecast quality as the probability of detection (POD) and the probability of false detection (POFD) (abbreviations according to Doswell et al., 1990). For other aspects of quality we chose to highlight the accuracy expressed with the proportion correct (PC) and the skill 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).

3 RESULTS

3.1 Rating of variables

The rating performed with 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 to define the target variable in comparison to the objective interpretation. Subsequently, we will only show results obtained with the subjective stability interpretation.

The two implemented stability indices Sk_{38} and *SSI* showed different discrimination power. For both – the detection of rather 'poor' and rather 'good' conditions – the Fisher criterion for Sk_{38} was larger than for *SSI*. In Fig. 1 the distributions of the two stability indices for the three observed stability categories are shown. While the Sk_{38} seems to be able to discriminate between the three categories (Kruskal-Wallis test, $p<0.001$), the *SSI* showed no significance ($p>0.05$). Also the combination of the indices was not significant (not shown).

For some of the variables which discriminated well between stable and unstable profiles for real snowpacks, i.e. failure layer grain size, hardness, and differences of these both to the adjacent layer (Schweizer and Jamieson, 2003) the sign of correlation was wrong. For example, while in unstable measured profiles the difference in grain size across the failure interface was large, it was small in modelled profiles.

For other significant variables in the real snowpack the sign of correlation agreed. More unstable profiles had shallower snow depth and lower failure layer shear strength. They were typically classified in profile types 7 and 4 (while more stable profiles were classified in profile type 6) (Schweizer and Lütschg, 2001). The failure layer in unstable profiles consisted more often of depth hoar (in the modelled stratigraphy also facetted crystals were correlated with more unstable profiles), while failure layers with rounded grains were found more often in stable profiles.

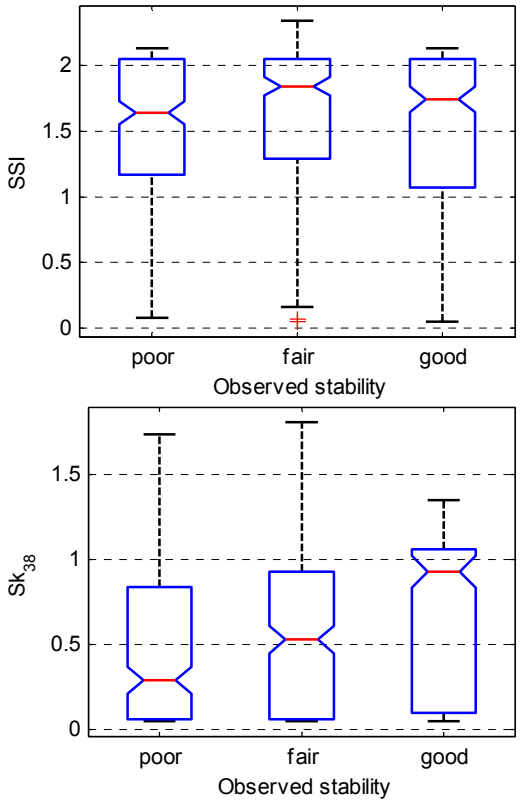

Fig. 1. Modelled stability indices vs. observed stability categories.

In contrast to the observations, the slab is 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 real snowpacks.

Other important variables, besides slab density, were mean slab properties such as hardness, grain size, bond size, and the product and ratio of grain size and bond size. Similar failure layer properties were rated as important as well. Furthermore, their differences between mean slab and failure layer properties showed high values of the Fisher criterion. No adjacent layer properties were chosen. Only a few meteorological variables were rated as important. For example, wind speed (24 hour mean), sensible heat fluxes (absolute difference between 24 hour maximum and minimum), and the 24 hour and 72 hour new snow sum. Furthermore, snow temperature at 10 cm below the surface, ski penetration depth and if a layer of depth hoar can be found one meter beneath the penetration depth. Variables representing a change in time were rarely rated as important. Most variables were important both for the detection of rather stable and rather unstable conditions.

For most variables which were previously 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 hours or the existence of a crust in the slab.

Many of the variables mentioned above were highly correlated. As input for classification trees only the best 20 not pair wise linearly correlated variables were considered (r^2<0.6). Variable selection was done for each cross-validation block (in this case: for each winter) separately, otherwise the performance will be overestimated.

3.2 Classification

The trees using the best 20 variables obtained with the Fisher criterion as input are shown in Fig. 2 for the detection of the rather stable conditions and in Fig. 3, for the detection of rather unstable conditions. The both trees 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 (not shown).

In Table 1 the cross-validated results for the detection of rather stable conditions are shown. In the row labelled median, performance measures were calculated on the test dataset in which each observation was set on the median value, when more than one measurement per day was available. In the row labelled orig., the original test dataset was used. Results for trees using only the input variable *SSI*, or the *SSI* and the Sk_{38} are not shown, since their performance measures were not as good as trees using only the Sk_{38}. The classification proposed by Schweizer et al. (2006) is also presented (labelled as 2006). Table 2 summarises the same results for the detection of the rather unstable conditions.

Best quality characteristics were achieved with the classification trees using the 20 best input variables. A true skill statistic (TSS) of 0.40 and 0.46 were reached for the detection of rather unstable and rather stable conditions respectively, combined with a proportion correct (PC) of 0.67 and 0.74. Also the Sk_{38} showed good results. Mostly no skill was obtained with the 2006 classification. The reason for this may 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.

Table 1. Overview of the cross-validated model results for the detection of rather stable conditions. Outlined are the proportion correct (PC), the probability of detection (POD), the probability of false detection (POFD) and the true skill statistic (TSS) for three different methods (described in the text). In the row 'median', median target values of each day were used, while in row 'orig.' the original observations were used for the test dataset. Base rate (fraction of observations of class 'good') was 0.3.

Model		PC	POD	POFD	TSS
best_20	orig.	0.68	0.64	0.30	0.34
	median	0.74	0.72	0.26	0.46
Sk38	orig.	0.59	0.76	0.50	0.26
	median	0.61	0.77	0.47	0.30
2006	orig.	0.62	0.30	0.24	0.06
	median	0.71	0.46	0.18	0.28

Table 2. Overview of the cross-validated model results for the detection of rather unstable conditions. Same abbreviations as in Table 1. Base rate (fraction of observations of class 'poor') was 0.3.

Model		PC	POD	POFD	TSS
best_20	orig.	0.63	0.68	0.39	0.29
	median	0.67	0.77	0.37	0.40
Sk38	orig.	0.63	0.62	0.37	0.25
	median	0.61	0.68	0.42	0.26
2006	orig.	0.57	0.44	0.37	0.07
	median	0.59	0.52	0.38	0.14

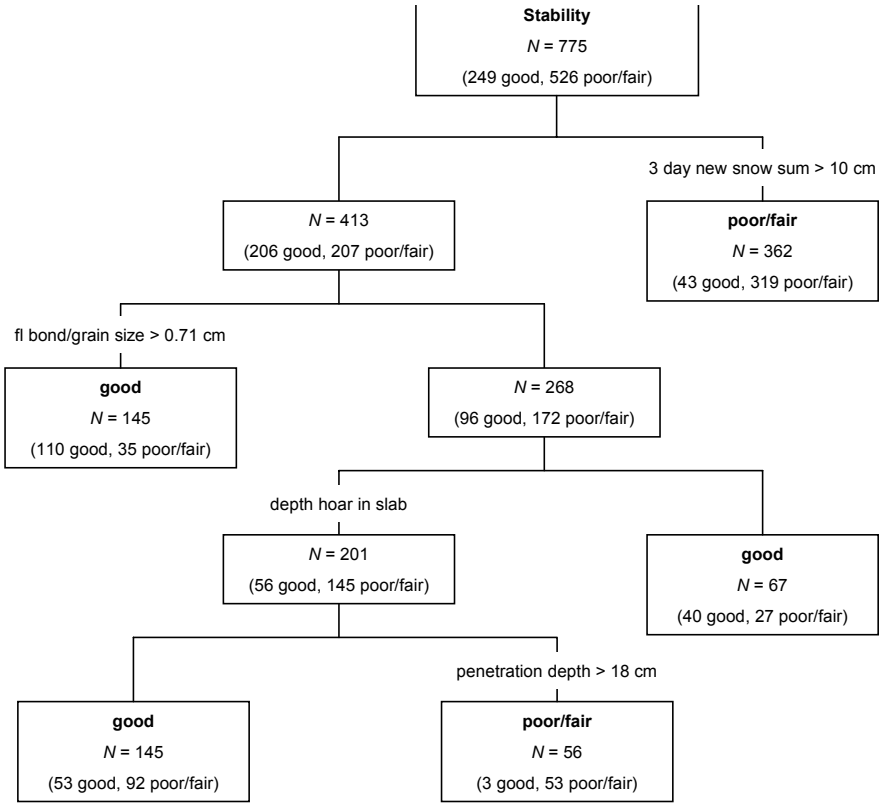

Fig. 2. Classification tree for the detection of rather stable conditions using the best 20 variables defined with the Fisher criterion.

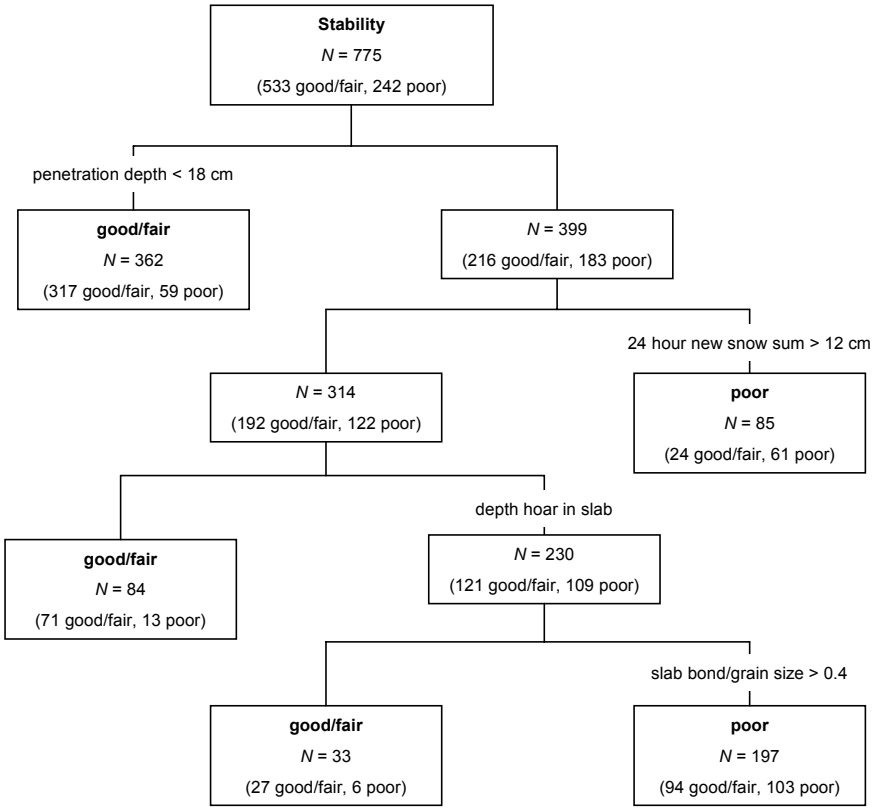

Fig. 3. Classification tree for the detection of rather unstable conditions using the best 20 variables defined with the Fisher criterion.

Better performance measures were gained for the detection of the rather stable conditions. This is consistent with larger values of the Fisher criterion. Also larger values were reached, when observations were rated with the median value of a given day. This may be explained with the in-region variance which cannot be produced by the models.

In a further step, the model stability estimation was verified only for time periods for which a very good estimate of the regional stability was available (Schweizer et al., 2003; Schweizer and Kronholm, 2007). Trees were built removing the whole corresponding winter for training. Results for the model *best_20* can be seen in Table 3, together with the verified and forecasted danger level. But what can be expected applying the two trees when compared to the 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 'Low', since the binary distribution 'good' vs. 'poor/fair' is 90% to 10%. Not as clear is the other situation: At danger level 'Considerable', the typical distribution of the binary variable 'poor' vs. 'fair/good' is mostly balanced. But since this tree produced more 'poor' situations than present in the observations (bias of 1.5 (Wilks, 1995)), we expect that the detection of rather unstable conditions might correspond to the detection of danger levels equal to and higher than 'Considerable'. The only day on which the models obviously failed is marked in grey in Table 3. For the other verified periods the models gave reasonable results.

Table 3. Results of the cross-validated classification trees using the best 20 variables as input (*best_20*) applied on periods with verified regional danger level.

Date	Forecasted danger level	Verified danger lelvel	Tree to detect 'poor'	Tree to detect '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	good
20 Mar 2002	3	3	poor	poor/fair
11-12 Dec 2002	2	1	poor	poor/fair
7 Jan 2003	2	3	poor	poor/fair
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

4 DISCUSSION AND CONCLUSION

The first objective of this paper 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} good values were obtained. The same conclusion can be drawn based on the Kruskal-Wallis test. This 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 real snow covers, have similar classification power (which was the second question of this paper). However, this hypothesis can now be rejected, at least for the four most important variables in real snow covers, namely difference in hardness and grain size across the failure interface, failure layer grain size and hardness, since an opposite sign of correlation to stability was found for modelled variables.

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 failure layer shear strength).

Of particular importance is that other variables, which cannot be measured or were not regarded yet (e.g. mean properties of the slab, differences of these to weak layer properties, and their change in time) do have good classification power. Most of these variables were also those that had been chosen independently by experts as important. The usefulness of modelled variables is further supported by the fact that the classification trees with the most important variables were able to distinguish "rather poor" or "rather good" conditions. Especially the performance for the ten periods with verified regional danger level is convincing.

When the target variable was defined with the subjective interpretation higher values of the Fisher criterion were obtained. This is surprising since the objective interpretation is consistent, while the subjective interpretation depends on the judgement of a person. Possibly, the advantage of the subjective interpretation, namely that the expert is able to include a broad spectrum of information in the rating, is balanced by the disadvantage of inconsistency. This explanation is supported by the fact that the modelled variables discriminated best between rather stable and rather unstable conditions when the subjective interpretation was applied.

Our models for stability evaluation can be used by avalanche warning services as nowcast or forecast in regions with similar climatic char-

acteristics as in the region of Davos. For regions with other characteristics, a new classification tree must be trained, which is only possible if a similarly large amount of observations is available as was the case in this study.

Nowcasting may have already an important value, since information on instability at the present day are frequently sparse. For forecasting, the present snow cover would be simulated with measured data, then the development of the snow cover would be predicted with forecasted meteorological data for the next day. The predicted snow cover finally 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.

In addition to other numerical forecasting methods that detect avalanche days or attempt to directly predict the avalanche danger our approach simply provides key data for the avalanche danger prediction process and is therefore particularly suited as a supporting tool for avalanche warning services.

ACKNOWLEDGEMENTS

This work has partially been funded by the Swiss National Science Foundation and the Swiss Federal Office of the Environment. 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 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 Reg. Sci. Technol. 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 Reg. Sci. Technol. 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 Forecast. 5, 576–585.

Durand, Y., Giraud, G., Brun, E., Mérindol, L., Martin, E., 1999. A computerbased system simulating snowpack structures as a tool for regional avalanche forecasting. J. Glaciol. 45 (151), 469–484.

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

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 Reg. Sci. Technol. 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 Reg. Sci. Technol. 35 (3), 169–184.

Lehning, M., Fierz, C., Brown, B., Jamieson, B., 2004. Modeling instability for the snow cover. Ann. Glaciol. 38, 337–338.

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

Schirmer, M., Lehning, M., Schweizer, J., 2009. Statistical forecasting of regional avalanche danger using simulated snow-cover. J. Glaciol. 55 (193), in press.

Schweizer, J., Bellaire, S., Fierz, C., Lehning, M., Pielmeier, C., 2006. Evaluating and improving the stability predictions of the snow cover model snowpack. Cold Reg. Sci. Technol. 46 (1), 52–59.

Schweizer, J., Föhn, P., 1996. Avalanche forecasting - an expert system approach. J. Glaciol. 42 (141), 318–332.

Schweizer, J., Jamieson, B., 2003. Snowpack properties for snow profile interpretation. Cold Reg. Sci. Technol. 37 (3), 233–241.

Schweizer, J., Kronholm, K., 2007. Snow cover spatial variability at multiple scales: Characteristics of a layer of buried surface hoar. Cold Reg. Sci. Technol. 47 (3), 207–223.

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

Schweizer, J., Lütschg, M., 2001. Characteristics of human-triggered avalanches. Cold Reg. Sci. Technol. 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 Reg. Sci. Technol. 51 (2-3), 112–121.

Schweizer, J., Wiesinger, T., 2001. Snow profile interpretation for stability evaluation. Cold Reg. Sci. Technol. 33 (2-3), 179–188.

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