Weather prediction using nearest-neighbour model

Dan Singh*, Ashwagosha Ganju and Amreek Singh

Research and Development Centre, Snow and Avalanche Study Establishment, Chandigarh 160 036, India

Prediction of any event using pattern recognition technique depends on the past history of the event. Past situations under similar conditions of the feature vector in feature space are used to predict the expected behaviour of the system. Based on this approach, a quantitative snowfall forecast model has been developed for a station in Jammu and Kashmir, using surface meteorological data of the past 12 winters (1991–92 to 2003–04, excluding the data of winter 1994–95, which was not available). The model predicts weather in terms of snow/no snow day and the amount of snowfall (snow height in cm) for three consecutive days in advance. The performance of the model has been tested for four winters for day-1, day-2 and day-3 forecasts. For qualitative snowfall forecast, the model performance for day-1, day-2 and day-3 forecasts turns out to be 80–90, 70–80 and 65–75%. The model estimates the expected snowfall amount at the station for day-1, day-2 and day-3 in advance, and based on the value of the estimated snowfall amount, it is categorized in the expected snowfall range based on the already established criterion. Quantitatively, the model predicts snowfall amount accurately for day-1 and the average accuracy of the model for different ranges of established categories varies from 25 to 55% for day-1 forecast. The model over-predicts the expected snowfall amount for day-2 and day-3 compared to day-1. The results of the models have been discussed here.

PRECIPITATION in North India during winter is mainly due to the development of lower latitudinal west and northwest air circulation, which gains moisture from the Arabian Sea and Persian Gulf, etc. during its eastward traverse. These westerly disturbances traverse through Afghanistan, Pakistan and finally approach Northwest Himalaya in the Indian region, where they precipitate snow (November to April).

Snow precipitation in this part of the country during winter makes the movement of civilions as well as army personnel difficult due to closure of roads/tracks and the development of hazardous situations due to avalanches. To make the best of fair-weather days, accurate weather and avalanche forecasts are required well in advance to ensure safe movement of personnel, necessary supplies and services along road axes/tracks. Thus there is need for a weather forecast model, which can predict weather in terms of snow/no snow days and likely amount of snow as well, in advance.

*For correspondence. (e-mail: dan_@rediffmail.com)

The approaching westerly disturbance causes certain changes in surface meteorological data, which are continuously measured and monitored over the North–Western Himalayan region, during winter. These perturbations in surface meteorological data are to some extent being tracked successfully by suitable numerical models and the future weather situations are inferred in advance. However, precise site-specific forecasts are not possible with these models.

Recently, many case-based reasoning techniques have been proposed to solve the problem of weather forecasting[1]. Intelligent systems (IS) using artificial intelligence techniques have been used to forecast visibility, marine fog, precipitation, severe weather and other climatological conditions[2–5]. The IS modelling approach is complementary to the NWP (Numerical Weather Prediction) scheme that uses computationally intense dynamical, thermodynamical and statistical algorithms to produce large-scale weather forecasts. These large-scale forecasts are not sufficiently stable on a small scale, where local effects become significant or even predominant. The large-scale numerical models can be scaled down using more detailed numerical models that take output from large-scale models and also include local effects and most recent data of weather stations in the neighbouring area. The smaller-scale numerical models are not intuitive, need consistent data and are based on climatological history of the area, where local effects are encountered in the data. The fuzzy case based system for weather prediction, which addresses the problem of forecasting horizontal visibility and cloud ceiling height at airport terminals has been proposed by Riordan and Hansen[1].

The model proposed here uses fuzzy similarity metric with built-in climatological knowledge. The case-based system retrieves stored cases from historical database using *k*-nearest-neighbour retrieval mechanism based on fuzzy similarity metric. Each retrieved case represents a previously encountered climatological situation that is similar to the current situation. The retrieved cases are adapted to construct a forecast scenario. Following the approach of case-based reasoning, the *k*-nearest neighbour model has been developed for predicting weather events qualitatively (snow/no snow day), and in a limited scope, the snow-fall amount in different qualitative categories. The model looks into the history of weather events for similar climatological situations and the outcome in terms of the predicate. The predicate outcome of the past situations is used in decision making for future weather outcome. The model developed so far has been tested during four winters. It generates three-day weather forecasts in advance in qualitative (snow/no snow day) and quantitative ways.

The nearest-neighbour model for prediction of avalanches was first proposed by Buser[6], which selects ten days most similar to the given situation from 20-years data. The model does not relieve the forecaster from making his own decision; it helps him prune down the decision. Supplying detailed information about similar situations in the past, the model can support local control in decisions to open ski runs or

whether to shoot or not, the critical avalanche slopes[7]. A more enhanced version of the nearest-neighbour model, NXD 2000, with optimizing weight and new elaborate variables, which brings knowledge of local avalanche forecaster into the model, has been proposed[8].

The nearest-neighbour model has the advantage of simple computational approach. The model represents only the most similar situations corresponding to the current situation and the decision making itself depends on the forecaster. Further, the selection of weights for different parameters is based on the intuition of the expert, although attempts have been made for optimizing weights also[8].

The Western Himalayan region comprises diverse climatic zones. Sharma and Ganju[9] have broadly classified the Western Himalaya into three major climatic zones, namely, Lower Himalayan Zone, Middle Himalayan Zone and Upper Himalayan Zone. The present study area lies in the Lower Himalayan Zone, i.e. Pir Panjal range of Northwest Himalaya, which is characterized by relatively warm temperature, heavy snow precipitation and short winter period. The westerly disturbance hits Pir Panjal range first in India. Due to high moisture content, heavy precipitation takes place in the Pir Panjal range and then it moves eastward, i.e. towards the Great Himalayan range and with little moisture content to Karakoram range. Snowfall starts early in November in Pir Panjal and frequent heavy snowstorms are dominant during mid-winter (February/March). The seasonal snow cover starts melting after mid-March in Pir Panjal and Great Himalaya. Precipitation thereafter is received as rain in Pir Panjal. Solid precipitation of the order of 50–60 cm is generally observed during normal snowstorm lasting 3 days and during extreme snowfall events, snowfall of the order of about 300 cm has been recorded during the past 12 winters. The mean ambient temperature varies from –5 to –10°C during peak winter and lowest temperature dips down to about –15°C. A brief description of meteorological conditions of the study area is presented in Table 1.

The Chowkibal–Tangdhar axis is the only road connecting the districts of Tithwal and Kupwara in J&K. The road axis has 26 major registered avalanche sites. The high-resolution meso-scale weather forecast models (e.g. MM5) are not meant for site-specific forecasts, as in the present case at Stage-II observatory, which is a representative weather station in the Chowkibal–Tangdhar area. The MM5 model used for forecast in the Chowkibal–Tangdhar area predicts weather parameters at 10 km resolution, which covers the whole area in 2 to 3 grid points (K. Srinivasan, pers. commun.). The NN model discussed here is likely to provide additional help to weather and avalanche forecasters.

The primary interest of this study is prediction of winter snowfall. Therefore, mostly data for the months from November to April are considered. The data are not strictly distributed during all winters from 1 November to 30 April. The final database considered here, is that of the past twelve winters (1991–92 to 2003–04) recorded at Stage-II observatory. The snow and weather data are continuously recorded by the winter study team as two daily measurements at 0830 UTC and at 1730 UTC. For the present study, data consisting of 4266 records of the period mentioned above, have been taken. Records for the surface meteorological parameters are not uniformly available for all winters due to the development of defects in the instruments, which could not be rectified during the course of a winter. Replacement/repairing of the instruments could not be undertaken, due to remoteness, inaccessibility and harsh and hazardous climatic conditions prevalent during winter. However, the records have been kept intact and the model has been run with missing parameters.

The basic concept of the nearest-neighbour model lies in the fact that similar situations will lead to similar outcomes. Thus the nearest-neighbour technique looks into the history of the events in the past data. The similarity of the present situation with the past ones is defined in terms of the similarity metric. For the development of the present model Euclidean metric has been taken as the similarity measure. The similarity metric between two days has been defined as:

Distance between day x_l and x_j

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2},$$

where x_l is the vector of m measurements for day i(m = 1, p); w_k is the weight factor for measurements.

The model algorithm, flow diagram and graphical representation of nearest-neighbour model in two dimensions are shown in Appendix 1.

The symbolic distances (representative of similarity of the present situation with past ones) of the present situation are calculated with the past situations, taking the actual outcome in terms of parameters taken for the decision-

Table 1. Climatic condition at Stage-II observatory

Month	Mean max. (°C)	Highest max. (°C)	Mean min. (°C)	Lowest min. (°C)	Mean cumulative snowfall (cm)	Mean standing snow (cm)
November	9.9	21.0	0.6	–8.0	51.0	20.45
December	5.5	15.0	–3.1	–10.0	88.0	28.8
January	2.5	12.0	–5.8	–16.0	268.5	78.1
February	3.0	12.5	–4.7	–12.5	287.2	122.6
March	6.1	16.0	–1.5	–14.0	214.0	142.9
April	13.3	22.5	3.5	–5.5	34.5	87.2

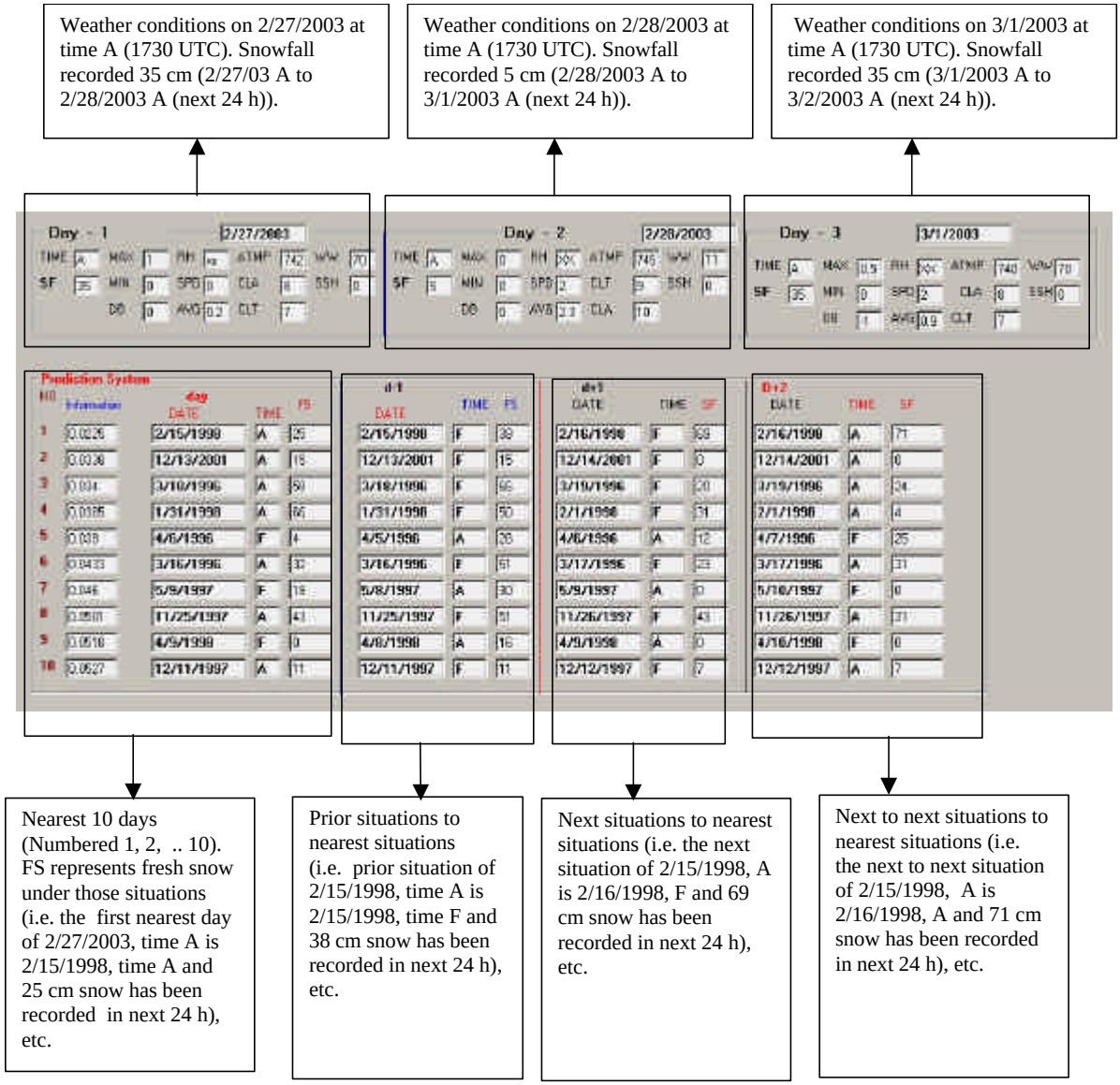


Figure 1. Model prediction scheme (nearest days/situations for 2/27/2003, time A (1730 UTC)).

making process. For the present case, the similarity of the present situation with the past has been calculated with equal weight to all parameters in the decision-making process. However, a few parameters may be biased. The ten nearest days of the past were selected for drawing the forecast scenario based on symbolic distances. The parameters selected for the present model development are given in Appendix 2.

The nearest ten days selected by the model are finally used in the decision criteria. The model is run for generating day-1, day-2 and day-3 forecasts, well in advance, based on the current and nearby situations. The model considers only the measurable snow precipitation in the past data; rainfall occurring in early and late winter is not considered. The problem lies in the fact that the wide volume of data representative of rainfall in the past years are not available.

The model run for 2/27/2003 at time A (1730 UTC) is shown in Figure 1. The snow and weather conditions for consecutive three days (2/27/2003 A, 2/28/2003 A and 3/1/2003 A) are shown and explained (Figure 1, top). The model finds the nearest days of the day 2/27/2003 A, and then analyses the data of the prior situation to the nearest situations, next day's situations to nearest situations and next to next day's situation to nearest situations (Figure 1, bottom). The trained model forecasts for three consecutive days, i.e. day-1 (2/27/2003 A to 2/28/2003 A), day-2 (2/28/2003 A to 3/1/2003 A), and day-3 (3/1/2003 A to 3/2/2003 A) under different categories (Table 2). The model forecast and observed weather for day-1, day-2 and day-3 are shown in Figure 2. The model forecast in terms of snow/no snow day is accurate for all three days and it failed in the expected snowfall category for day-2 forecast (Figure 2).

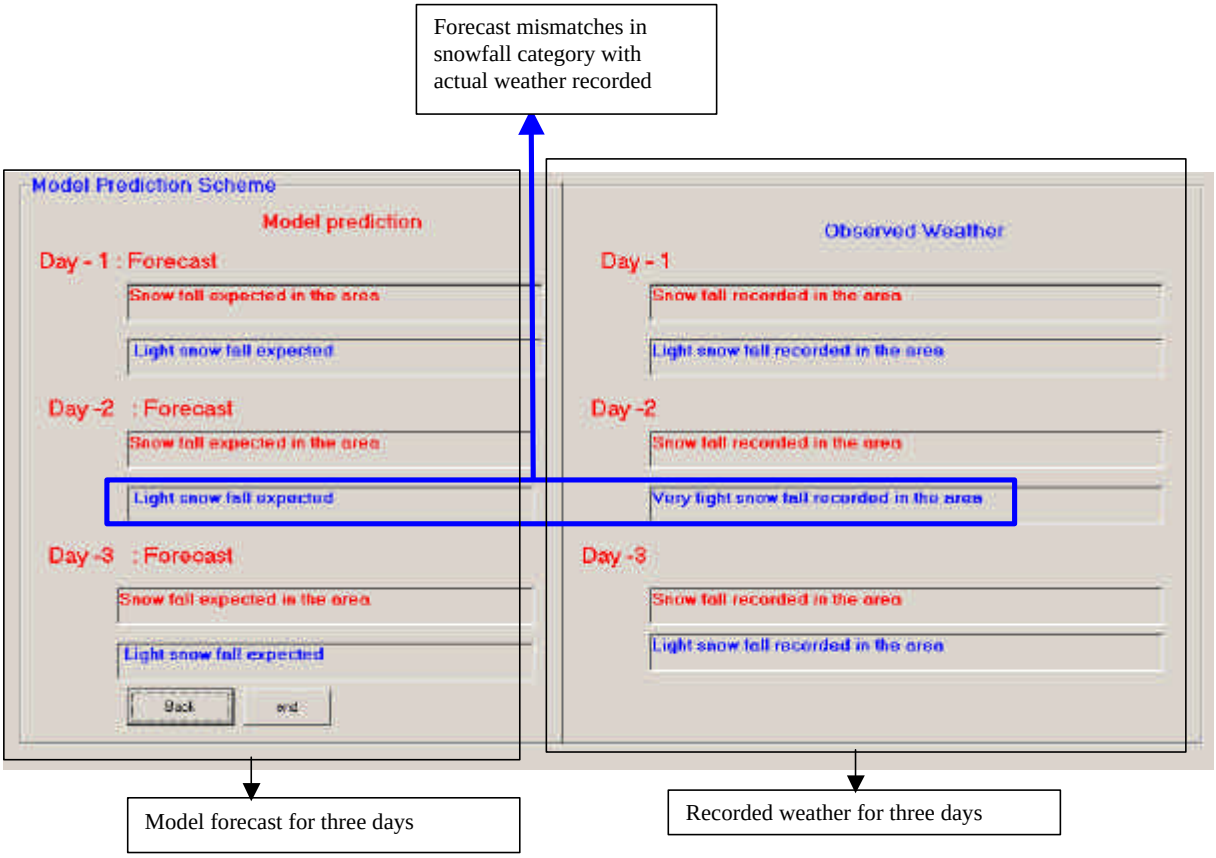


Figure 2. Final weather forecast generated by the model.

Table 2. Categorized snowfall ranges

Name	Category	Range
A	Very light snowfall	≤ 12 cm
B	Light snowfall	12.1–35.0 cm
C	Moderate snowfall	35.1–65.0 cm
D	Heavy snowfall	> 65 cm

The method has the advantage of simple computational approach for prediction of weather compared to the complex computations as well as simulation of real weather events in dynamical models. The method is not useful for long-range weather prediction (for more than three days) due to the effect of approaching westerly disturbances on the surface meteorological parameters. For the present study, the ten nearest neighbours have been selected. As the data volume increases in near future, the model performance will improve and stabilize.

The developed model was trained with the data of winter 1997–98, taken as a normal winter where snowfall is more or less uniformly distributed. The basic purpose of training the model is to arrive at the threshold value, which would determine whether a predicted day could be taken as a snowfall day or not, based on the probability score achieved in the model. The decision criterion of the model for day-1, day-2 and day-3 forecasts was decided in terms of probability of snowfall, which comes out to be 40, 40 and 35% respectively. The criterion thus developed was tested with the data of four winters (1998–99, 1999–2000, 2001–02 and 2002–03) for generating qualitative and quantitative snowfall forecast for day-1, day-2 and day-3. The model first predicts whether the predicted three days could be snow/no snow days. If the day has been predicted as snow day, then it predicts the expected snowfall amount for day-1, day-2 and day-3 based on criteria explained in succeeding paragraphs. The qualitative prediction of the model for day-1, day-2, and day-3 lies between 80 and 90%, 70 and 80%, and 65 and 75% (Table 3) for test winters. The qualitative forecast of the model decreases as we proceed from day-1 to day-2 and day-3. The overall prediction scheme of the model is given in Figures 1 and 2.

The model calculates the average value of snowfall in the nearest ten days, ten next situations to nearest situations and ten next to next situations to nearest situations at the station for estimation of expected snowfall amount for day-1, day-2 and day-3 forecasts. Based on the average value of snowfall amount, the model forecasts are divided into four groups, A, B, C, and D (Table 2). This classification of snowfall into different categories is used for weather

Table 3. Model performance (qualitative snowfall forecast)

	Day-1		Day-2		Day-3	
	SF days	NSF days	SF days	NSF days	SF days	NSF days
Winter 1998–99						
SF days	32	9	23	19	21	21
NSF days	10	126	19	116	27	107
Accuracy	88.76%	78.53%	72.31%			
Winter 1999–2000						
SF days	26	13	18	21	17	22
NSF days	14	129	28	114	44	97
Accuracy	85.16%	72.92%	63.33%			
Winter 2001–02						
SF days	25	7	14	18	17	15
NSF days	23	126	33	115	39	108
Accuracy	83.42%	71.67%	69.83%			
Winter 2002–03						
SF days	23	11	13	21	14	20
NSF days	13	134	31	115	28	117
Accuracy	86.74%	71.11%	73.18%			

SF, Snowfall; NSF, No snowfall.

Table 4. Model performance (quantitative forecast)

	Day-1						Day-2						Day-3					
Class	Days	A	B	C	D	E (%)	Days	A	B	C	D	E (%)	Days	A	B	C	D	E (%)
Winter 1998–99																		
A	7	3	4	–	–	42.8	9	1	7	1	–	11.1	9	3	5	1	–	33.3
B	23	7	12	4	–	52.17	8	2	4	2	–	50	12	3	9	–	–	75
C	2	1	1	–	–	0	6	1	5	–	–	0	–	–	–	–	–	–
D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Winter 1999–2000																		
A	7	4	2	1	–	57.1	10	2	4	–	4	20	11	6	5	–	–	54.6
B	19	6	8	2	3	42.1	6	3	2	1	–	33.3	6	2	2	2	–	33.3
C	–	–	–	–	–	–	2	–	–	1	1	50	–	–	–	–	–	–
D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Winter 2001–02																		
A	4	2	2	–	–	50	6	2	2	–	2	33.3	11	5	4	1	1	45.6
B	16	3	6	4	3	37.5	7	2	2	1	2	28.6	6	2	2	2	–	33.3
C	5	1	2	1	1	20	1	–	–	1	–	100	–	–	–	–	–	–
D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Winter 2002–03																		
A	9	6	2	1	–	66.7	3	–	3	–	–	0	5	4	–	1	–	80
B	13	1	8	2	2	61.5	7	2	1	2	2	14.3	9	4	1	2	2	11.2
C	1	–	–	–	1	0	3	1	–	1	1	100	–	–	–	–	–	–
D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

forecasting in avalanche-prone areas of India in terms of rain (in mm). The rain (in mm) has been converted into equivalent snow amount (in cm) assuming average snow density equal to 100 kg/m^3.

The accuracy of the model is high for day-1 forecast, with few cases of under-prediction (Table 4). For day-2 and day-3 forecasts, the accuracy of the model decreases and over-prediction of the model increases in different categories. Overall, the model prediction for most occurring snowfall events in Chowkibal–Tangdhar axis is high compared to extreme snowing events.

The accuracy of the model for prediction of heavy snowfall events for day-1 forecast is less; the cases are either under-predicted or over-predicted by the model. The quantitative prediction of the model is given in Table 4.

The model predicted 7 days under category A, 23 days under category B, 2 days under category C and no day under category D for day-1 forecast for winter 1998–99

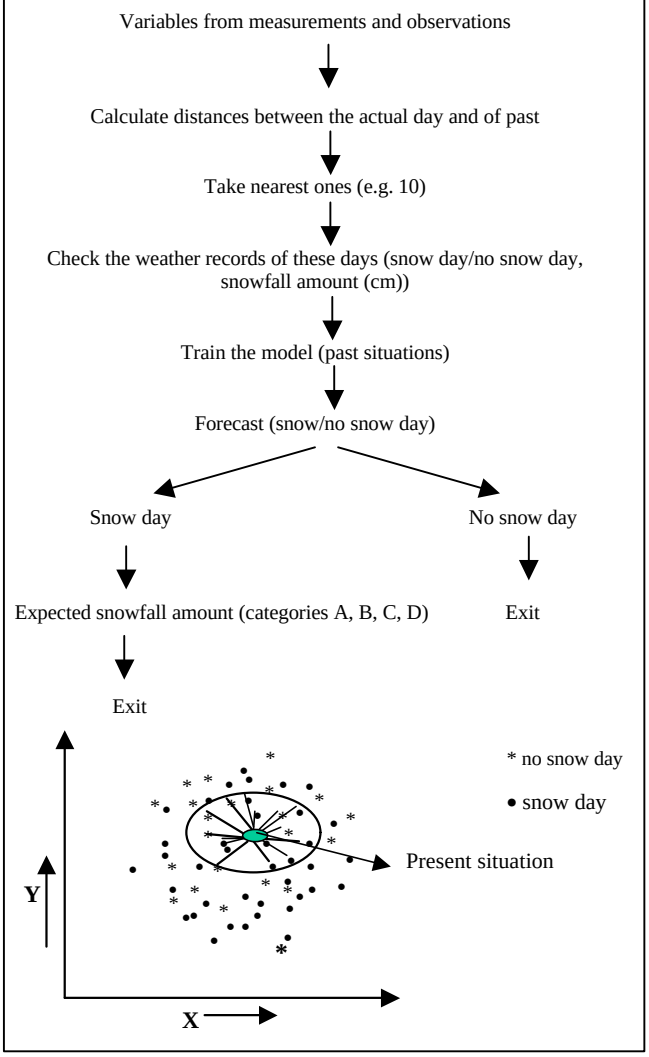
(Table 4). Out of 7 days under category A, snowfall was recorded for 3 days (accuracy 42.8%) under category A and out of 23 days under category B, snowfall was recorded for 12 days (accuracy 52.17%) under category B, etc. The accuracy of the model is given under column E (Table 4) for different categories. The model performance has been calculated in a similar way for day-2 and day-3 forecasts and for different winters. Extension of the model for more accurate quantitative as well as qualitative prediction is under progress.

The performance of the model developed so far is satisfactory with respect to the available information. Data for all parameters are not available for a few winters due to sudden failure of meteorological instruments. Once all the parameters are available in the database in future, the model performance is likely to improve. To overcome the problem of missing data, an automated recording of a few weather parameters and installation of state-of-the-art equipment have been initiated.

Further, in the present model all parameters have been given equal importance in the decision-making process. Biasing of the parameters, which is being attempted in terms of the dynamic weights (computed from the data at each run), may provide better results. The model will be helpful to regulate the mobility of personnel along Chowkibal–Tangdhar axis, so that they do not get trapped during heavy snowfall. This model will also provide additional help to the avalanche forecaster for assessing avalanche danger well in advance along the road axis. Different weights may be assigned to different days for computing expected snowfall amount for better prediction.

The proposed model predicts weather at a representative station on the Chowkibal–Tangdhar axis. This can be extended in the remaining areas of the Northwest Himalaya in J&K, which may provide the overall picture of the snow precipitation at different places during a winter.

Appendix 1. Model algorithm, flow diagram and graphical representation of nearest neighbour model



Appendix 2. List of parameters used for model development and their abbreviation.

Date
Time (F (0830 UTC), A (1730 UTC))
Maximum temperature (MAX)
Minimum temperature (MIN)
Ambient temperature (DB)
Relative humidity (RH)
Cloud amount (CLA)
Cloud type (CLT)
Spot wind speed (SPD)
Average wind speed (AVG)
Atmospheric pressure (ATMP)
Weather code (WW)
Sunshine hours (SSH)

1. Riordan, D. and Hansen, K. B., A fuzzy case based system for weather prediction. *Eng. Int. Syst.*, 2002, **10**, 139–145.
2. Conway, B. J., Expert system and weather forecasting. *Meteorol. Mag.*, 1989, **118**, 23–30.
3. Hall, T., Precipitation forecasting using neural network. *Weather Forecast.*, 1999, **14**, 338–346.
4. Marzban, C. and Stunpf, G. J., A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteorol.*, 1996, **35**, 617–626.
5. Marzban, C., A Bayesian network for severe hail prediction. *Weather Forecast.*, 2001, **16**, 600–661.
6. Buser, O., Avalanche forecast with the method of nearest neighbours: An interactive approach. *Cold Regions Sci. Technol.*, 1983, **8**, 115–163.
7. Buser, O., Monika, B. and Walter, G., Avalanche forecast by nearest neighbour method. IAHS Pub. No. 162, 1987, pp. 557–567.
8. Martin, G., Hans, J. E., Karl, B. and Tom, L., NXD 2000 – An improved avalanche forecasting programme based on nearest neighbour method. ISSW, 2000, pp. 52–59,
9. Sharma and Ganju, Complexities of avalanche forecasting in Western Himalaya: An overview. *Cold Region Sci. Technol.*, 2000, **31**, 95–102.
10. Battan, L. J., *Fundamentals of Meteorology*, Prentice-Hall, New Jersey, 1984, pp. 157–161.
11. Buser, O., Fohn, P., Gubler, H. and Salm, B., Different methods for assessment of avalanche danger. *Cold Regions Sci. Technol.*, 1985, **10**, 199–218.

12. Buser, O., Two years' experience of operational avalanche forecasting using nearest neighbours method. *Ann. Glaciol.*, 1989, **13**, 31–34.
13. Carter, M. M. and Elsner, J. B., A statistical method forecasting rain over Puerto Rico. *Weather Forecast.*, 515–525.
14. Fukunaga, K. and Hostetler, Optimization of *k*-nearest neighbour density estimate. *IEEE Trans. Inf. Theor*., 1973.
15. Thomas, J. H., Xufeng, N. and James, E. B., A space–time model for seasonal hurricane prediction. *Int. J. Climatol*., 2002, **22**, 451–465.
16. Pirmin, K. and David, W. C., Cluster-analysis classification of winter wind patterns in the Grand Canyon region. *J. Appl. Meteorol.*, 1999, 1131–1147.
17. Mohanty, U. C. and Dimri, A. P., Location-specific prediction of the probability of occurrence and quantity of precipitation over western Himalaya. *Weather Forecast*., 2004, **19**, 36–49.
18. Wilks, D. S., *Statistical Methods in Atmospheric Sciences*, Academic Press, New York, 1995.

ACKNOWLEDGEMENTS. We thank Dr R. N Sarwade, Director, Snow and Avalanche Study Establishment (SASE), Chandigarh for encouragement and scientists and technical assistants of Avalanche Forecasting Group, SASE who collected data under harsh climatic conditions in remote snow-bound areas.

Received 13 August 2004; revised accepted 7 January 2005

Synthetic accelerograms for two Himalayan earthquakes using convolution

V. N. Singh[1] and Abha Mittal[2,*]

[1]Department of Earth Sciences, Indian Institute of Technology, Roorkee 247 667, India
[2]Central Building Research Institute, Roorkee 247 667, India

In the present communication, computation of synthetic accelerograms is based on convolution. The spectrum of ground motion expected at a recording site is first computed from a knowledge of source parameters and medium properties. This spectrum is then inverse Fourier transformed to yield the desired synthetic accelerogram. This method has been successfully used by Boore, and has been further extended in the present communication. The suitability of the method is demonstrated successfully by modelling the accelerograms for two Himalayan earthquakes namely, the 1991 Uttarkashi earthquake and the 1999 Chamoli earthquake and compared with the observed accelerograms.

EARTHQUAKE-resistant design of engineering structures is one of the most important methods of mitigating risk of damage from future earthquakes. Such designs are based on the specification of ground motion which can be expected in the event of an earthquake. However, for earthquake-resistant design of some important structures like dams and nuclear power plants, located in seismically active areas, it is desirable to have a reliable site-specific design accelerogram. Available records of strong ground motion, after suitable modifications, have been used in the past for detailed dynamic analysis of engineering structures. However, synthetic accelerograms are now increasingly being used in earthquake engineering. A knowledge of regional and local seismicity and seismotectonics, a suitable earth model and source characteristics of the design earthquake are required for this purpose.

There have been some recent attempts on new approaches to synthesize strong ground motions and to obtain source parameters[1–12].

Khattri *et al.*[13] and Yu *et al.*[11] carried out synthesis of strong motion for the Uttarkashi earthquake. Khattri *et al*.[13] carried out forward modelling using the isochrone method and inverted the observed accelerograms by recursive stochastic inverse algorithm to obtain the earthquake source slip function. Yu *et al.*[11] generated synthetics using the composite source model and synthetic Green's function. Solution of the forward problem carried out by Yu *et al*.[11] has taken into account the velocity structure in the Uttarkashi area and its *Q*-structure. The method of generating synthetics depends on the knowledge of many input parameters like velocity and *Q*-structure of the layered earth model. Kumar *et al.*[12] used semi-empirical method for calculating synthetic accelerograms. They divided the fault plane into sub-faults and generated envelop waveform, instead of actual time history, corresponding to each element of the fault plane.

In the present communication, an improved method of generating synthetic accelerograms has been presented and discussed. Generation of synthetic accelerograms in the near field is based on a dislocation moving over a fault plane. The computed ground motions have to take into account the nature of rupture propagation over the fault plane, radiation pattern effects, presence of free surface layering in the earth between the source and free surface and effect of finite moving source. The slip on the causative fault is specified in terms of shape, rise time and amplitude of the source time function. In addition, velocity of rupture propagation and final area over which slip occurs are also specified.

Let $y(t)$ represent the recorded seismogram at a point on the surface of a layered half space produced by a point shear dislocation. This can be written as:

$$y(t) = C*s(t)*a(t)*d(t)*i(t), \quad (1)$$

where C is a scalar, $s(t)$ is the source time function, $a(t)$ represents the impulse response of the layered medium between source and receiver, $d(t)$ accounts for frequency-dependent attenuation and $i(t)$ is the impulse response of the seismograph. In frequency domain, eq. (1) can be written as:

$$Y(w) = CS(w)A(w)D(w)I(w), \quad (2)$$

*For correspondence. (e-mail: abham2003@yahoo.com)