J Arid Land
https://doi.org/10.1007/s40333-018-0110-2

Science Press Springer-Verlag

Environmental factors influencing snowfall and snowfall prediction in the Tianshan Mountains, Northwest China

ZHANG Xueting[1,2], LI Xuemei[1,2*], LI Lanhai[3], ZHANG Shan[1,2], QIN Qirui[1,2]

[1] Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China;
[2] Gansu Provincial Engineering Laboratory for National Geographic State Monitoring, Lanzhou 730070, China;
[3] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China

Abstract: Snowfall is one of the dominant water resources in the mountainous regions and is closely related to the development of the local ecosystem and economy. Snowfall predication plays a critical role in understanding hydrological processes and forecasting natural disasters in the Tianshan Mountains, where meteorological stations are limited. Based on climatic, geographical and topographic variables at 27 meteorological stations during the cold season (October to April) from 1980 to 2015 in the Tianshan Mountains located in Xinjiang of Northwest China, we explored the potential influence of these variables on snowfall and predicted snowfall using two methods: multiple linear regression (MLR) model (a conventional measuring method) and random forest (RF) model (a non-parametric and non-linear machine learning algorithm). We identified the primary influencing factors of snowfall by ranking the importance of eight selected predictor variables based on the relative contribution of each variable in the two models. Model simulations were compared using different performance indices and the results showed that the RF model performed better than the MLR model, with a much higher R^2 value (R^2=0.74; R^2, coefficient of determination) and a lower bias error (RSR=0.51; RSR, the ratio of root mean square error to standard deviation of observed dataset). This indicates that the non-linear trend is more applicable for explaining the relationship between the selected predictor variables and snowfall. Relative humidity, temperature and longitude were identified as three of the most important variables influencing snowfall and snowfall prediction in both models, while elevation, aspect and latitude were of secondary importance, followed by slope and wind speed. These results will be beneficial to understand hydrological modeling and improve management and prediction of water resources in the Tianshan Mountains.

Keywords: snowfall prediction; snowfall fraction; random forest; multiple linear regression; predictor variables; Tianshan Mountains

Citation: ZHANG Xueting, LI Xuemei, LI Lanhai, ZHANG Shan, QIN Qirui. 2019. Environmental factors influencing snowfall and snowfall prediction in the Tianshan Mountains, Northwest China. Journal of Arid Land, 11(1): 15–28. https://doi.org/10.1007/s40333-018-0110-2

1 Introduction

The arid and semi-arid regions in Northwest China, located in the hinterland of the Eurasian continent in the mid-latitudinal zone, are sensitive to global climate change (Chen et al., 2014). Climate in Northwest China has gradually changed from warm-dry to warm-wet since 1987 (Shi et al., 2007). Both temperature and precipitation have shown increasing trends in Northwest

*Corresponding author: LI Xuemei (E-mail: shuimingren@163.com)
Received 2017-12-20; revised 2018-11-15; accepted 2018-12-03

China, particularly in the mountainous regions within the last 50 years (Li et al., 2013). Results from Füssel and Jol (2012) suggest that variations in temperature and precipitation within a wide range of mountainous regions are much greater than the global average. Snowfall, the major component of solid precipitation and the defining constituent in the mountainous regions, is regarded as a crucial index of climate fluctuations coupled with snow cover (Piazza et al., 2014; Mir et al., 2015). Winter snowfall often accumulates on the ground and translates into snow cover, which increases surface albedo and impacts surface runoff and energy budget (Dai, 2008).

Mountainous regions are often characterized by complex topography and diverse landscapes, with various forms of precipitation occurring at different elevation gradients and seasons. For example, rain falls in warm areas at lower elevations while snowfall often occurs in cold areas at higher elevations (Marks et al., 2013). The amount of snowfall is extremely important in the mountainous regions because of the process of snow accumulation and melt (Zhang et al., 2012; Krasting et al., 2013; Yu et al., 2015). Snowfall is also closely related to local tourism and has attracted research interests in natural disasters, such as avalanches (Scipiõn et al., 2013; Nair et al., 2017; Vrotsou et al., 2017).

In Xinjiang of China, the Tianshan Mountains are typically referred to as a "solid reservoir". Precipitation primarily occurs in the form of snowfall in winter, which accounts for more than 30% of the total annual precipitation in the western Tianshan Mountains (Lu et al., 2016). Alpine regions also have snowfall in summer (Shen et al., 2016). Approximately 373 rivers originate from the Tianshan Mountains and the variations in runoff are closely related to climate change (Xu et al., 2014; Wang et al., 2016). The Tianshan Mountains have an annual runoff of 47.4×10^9 m^3, accounting for 54% of the total river runoff from Xinjiang (Hu, 2004). Moreover, the Tianshan Mountains have some of the most extensive modern midlatitude glaciers, which are located at an altitude of 3500 m a.s.l. despite continuing glacial melt due to global warming (Sorg et al., 2012; Chen et al., 2017). Glacial meltwater contributes to the stable base flow for rivers in the alpine basins (Zhang et al., 2016). Water resources from snowmelt and glacial meltwater in the Tianshan Mountains play a key role in the continental hydrologic cycle, agriculture and industry development in Xinjiang.

Previous research in the Tianshan Mountains primarily focused on describing the temporal and spatial variations in snowfall quantity and snow cover, such as days of snow cover and snow-covered area (Xu and Qiu, 1996; Li et al., 2012; Liu et al., 2012; Guo and Li, 2015; Zhang et al., 2015; Chen et al., 2016, 2017; Li et al., 2016; Jing et al., 2017; Tang et al., 2017). The northern Tianshan Mountains are the principal area of snowfall in China when considering snowfall quantity and day metrics (Liu et al., 2012; Zhang et al., 2015). Winter snowfall in the Tianshan Mountains has exhibited a significant increasing trend and cyclical variations since 1961, and showed a clear heterogeneity at the spatial scale (Xu and Qiu, 1996; Li et al., 2012). Guo and Li (2015) and Chen et al. (2016) found that the ratio of snowfall to precipitation (snowfall fraction) has decreased in parts of the middle Tianshan Mountains in recent years, accompanied by increases in temperature. Jing et al. (2017) applied a general circulation model (GCM) to investigate future changes in snowfall and precipitation in the Tianshan Mountains and north of the Kunlun Mountains. They proposed that both the average annual snowfall and snowfall fraction will decrease significantly by the end of the 21st century.

The numbers of meteorological stations are limited in the mountainous regions, especially at higher elevations with abundant snowfall. The traditional approach for estimating snowfall based on meteorological data assumes that the relationship between snowfall and elevation is constant (Clark and Andrew, 2006). However, this approach has been proven inaccurate due to the scarcity of meteorological stations and the spatial variability of snowfall in the mountainous regions (Asaoka and Kominami, 2012). Asaoka and Kominami (2012) reconstructed the distribution of snowfall in the mountainous regions of Japan using satellite observations, and Scipiõn et al. (2013) obtained estimates of alpine snowfall amount using radar application. Yet, forecasts of meteorological elements were inaccurate because radar coverage in the mountainous regions are sparse due to the influence of topographic blocking (Wetzel et al., 2004; Clark and Andrew, 2006). Global and regional models have both overestimated and underestimated annual snowfall at

different elevation levels; however, a weather research and forecasting (WRF) model could perform well in an area with complex topography, although it requires a high resolution (Ikeda et al., 2010; Rasmussen et al., 2011).

In the mountainous regions, moist airflow is forced upward along the windward slope because of orographic effects, leading to increased localized precipitation. In addition, the complex terrain and landform of the mountainous regions have been well-characterized. Therefore, the influence of topography on precipitation (quantity and type) cannot be ignored in the mountainous regions. Recent studies have established linear relationships of precipitation with geographical and topographic variables, which were then used to estimate precipitation in the Tianshan Mountains (Ji and Chen, 2012; Zhang et al., 2015). A similar linear function was also used to model precipitation with three independent variables, including longitude, latitude and elevation, in the Appalachian Mountains (Padoan et al., 2009). Some studies parameterized only a single or a few variables of topography and climate to detect their impact on snowfall (Karl and Groisman, 1993; Davis et al., 1999; Perry and Konrad, 2006; Asaoka and Kominami, 2012). In general, most studies have focused on precipitation estimates or used a physical model to simulate snowfall with various model resolutions. There is a lack of multivariate coupled and non-linear methods to estimate snowfall accurately, especially in the mountainous regions where meteorological observations are scarce. Thus, it is critical to explore the main driving factors on snowfall to improve its estimation.

In this study, we introduced a relatively novel machine learning algorithm to identify the primary variables influencing snowfall using the level of importance of the selected variables in the random forest (RF) model and multiple linear regression (MLR) model. The proposed RF model has been widely applied to ecological modeling (Muñoz and Felicísimo, 2004; Cutler, 2007). Tinkham et al. (2014) recently used the model to determine the distribution of snow depth in a mountain catchment.

The objectives of this study were to (1) evaluate the main controlling factors of snowfall in the Tianshan Mountains in Xinjiang by identifying and ranking the order of independent variables, and (2) predict snowfall using the RF and MLR models. We further compared the performance of each model with the measured data. Accurate snowfall predictions are critical to better understanding surface hydrological cycle processes and parameterizing hydrological models and the spatial variability of snowfall distribution in the mountainous regions with complex topography. Furthermore, accurate modeling will also be a valuable input in assessments of water resources, local ecosystems and economies in the mountainous regions.

2 Materials and methods

2.1 Study area

The Tianshan Mountains in Xinjiang of Northwest China extend 1700 km from east to west with a width of 250–300 km, which account for more than 34.5% of the total area of Xinjiang (Hu, 2004) (Fig. 1). Far from the ocean, the mountainous regions are predominantly characterized by temperate continental arid climate. Geographically, the Tianshan Mountains separate the Xinjiang into northern and southern parts, resulting in regional climate differences. Influenced by the westerlies and topography, precipitation in the Tianshan Mountains shows uneven spatial and temporal distribution, with precipitation higher on the northern slope than the southern slope. The Tianshan Mountains are composed of a series of mountains, basins and plains. Less precipitation occurs in the intermountain basins and more precipitation in the mountain regions. The average elevation of the ridgeline is 4000 m a.s.l. in the Tianshan Mountains, and the temperature is particularly affected by the elevation. Annual average temperature in the Tianshan Mountains varies spatially, lower on the northern slope while higher on the southern slope, and lower in the mountains while higher in the plains.

2.2 Data collection

Generally, daily snowfall is recorded as liquid water equivalent when snow occurs. Mountainous

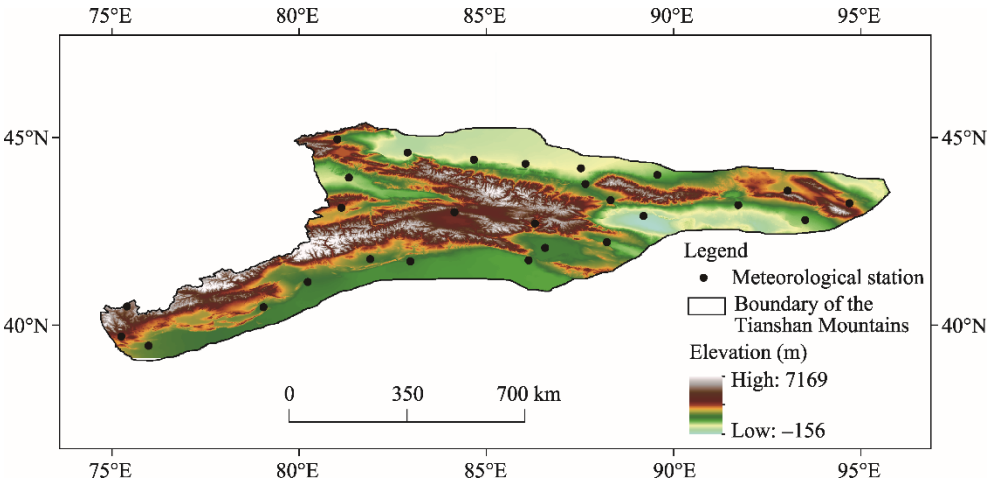

Fig. 1 Overview of the Tianshan Mountains and distribution of the 27 meteorological stations used in this study

precipitation is measured using weighing gauges, which have not distinguished between solid and liquid precipitation since 1980. Thus, we obtained daily snowfall data by separating solid and liquid precipitation based on temperature calculated by Zhang et al. (2017) from 27 meteorological stations in the Tianshan Mountains. The study period was defined as October to April (cold season) during the period from 1980 to 2015 because 99% of snowfall occurs between October and April annually (Guo and Li, 2015). The variables affecting snowfall mainly include climatic, geographical and topographic records. Climate data that included daily temperature, relative humidity and wind speed were derived from the National Meteorological Information Center in China (http://data.cma.cn). Topographic factors (including elevation, slope and aspect) were obtained from a digital elevation model (DEM) at a 30-m resolution provided by the Geospatial Data Cloud in China (http://www.gscloud.cn). A double mass curve method was applied to achieve data quality control by detecting outliers. The RF model is robust even with observations with missing and noise features. The prediction of the response variable can be implemented even with partially built trees; therefore, it is unnecessary to preprocess any missing data before inputting (Antipov and Pokryshevskaya, 2012).

2.3 Prediction models

In this study, we used the MLR model and RF model to predict the amount of snowfall by applying several predictor variables and response variable. The MLR model was used as a linear statistical method to explain the variance of dependent variable and multiple independent variables and to explore the degree of correlation among these variables. The MLR model was implemented using the *lm* function in R software. The metrics *lmg* was employed to evaluate the contribution of each predictor variable to the model because it measured the relative contribution of each variable to the coefficient of determination. The *lmg* was also executed in R software using the Relaimpo package (Grömping, 2006; Kousari et al., 2011; Oliveira et al., 2012).

RF model, a non-linear and non-parametric machine learning algorithm proposed by Breiman (2001), is assembled with a variety of decision trees generated from classification and regression trees (CART). The regression technique was selected to make predictions combining with multiple regression trees. Each regression tree was built by bootstrap samples, which were from the whole calibration datasets by randomly selecting a characteristic subset of the predictor variables to split at each node (Genuer et al., 2010). The corresponding output is the average of the individual regression tree. It is less susceptible to the problem of overfitting and no pruning step is required with the improved prediction accuracy compared with CART (Antipov and Pokryshevskaya, 2012). A prominent feature of the RF model is that it leaves out one third of the samples from the original sets as out-of-bag (OOB) samples to validate forecast accuracy (Oliveira et al., 2012). OOB samples were not used for fitting the regression tree in the process of

the RF model construction but were used to calculate the importance of predictor variables and estimate the prediction error. In other words, unbiased estimation of prediction error was built using OOB samples without any other independent test sets or cross-validation.

There is a criteria to measure the variable importance on the basis of the mean decrease in accuracy (%IncMSE) in the RF model, which is more reliable than the Gini importance, a biased measure of impurity decrease (Strobl et al., 2008; Genuer et al., 2010). The RF model was implemented using the Random Forest package in R software. To run the model, it is essential to define three prior parameters (*ntree*, the number of trees to grow in the model; *mtry*, the number of variables randomly sampled as candidates at each split; and *nodesize*, the minimum size of terminal nodes) in the forest. An optimal *ntree* value was set as 500 (see Palmer et al. (2007) for a detailed explanation). The value of *mtry* was set to 3 for 8 predictor variables (the number of predictor variables divided by 3 for the regression). Increasing or decreasing the *mtry* value had very little impact on model performance while a small change was found in the ranking of variable importance. Oliveira et al. (2012) also found that the *mtry* value had an insignificant effect on the sequence of variable importance. The value of *nodesize* was set to 5 (default value) when building regression trees, namely the minimum number of the leaf. A larger *nodesize* would cause smaller trees to grow but take less time.

2.4 Model performance

The dataset was randomly divided into two independent sections: 70% of data for calibration and remaining 30% of data for validation (period from October 1980 to April 2015). The evaluation indices of model performance for calibration and validation are as follows: coefficient of determination (R^2), index of agreement (d) and the ratio of root mean square error to standard deviation of observed dataset (*RSR*). They are described as:

$$R^2 = \frac{\left[\sum_{i=1}^{N}\left(O_i-\overline{O}\right)\left(P_i-\overline{P}\right)\right]^2}{\sum_{i=1}^{N}\left(O_i-\overline{O}\right)^2\sum_{i=1}^{N}\left(P_i-\overline{P}\right)^2}, \tag{1}$$

$$d = 1.0 - \left[\frac{\sum_{i=1}^{N}\left(P_i-O_i\right)^2}{\sum_{i=1}^{N}\left(\left|P_i-\overline{O}\right|+\left|O_i-\overline{O}\right|\right)^2}\right], \tag{2}$$

$$RSR = \frac{\left[\sqrt{\sum_{i=1}^{N}\left(O_i-P_i\right)^2}\right]}{\left[\sqrt{\sum_{i=1}^{N}\left(O_i-\overline{O}\right)^2}\right]}, \tag{3}$$

where O_i and P_i are the observed and predicted values, respectively; $\overline{O}$ and $\overline{P}$ are the mean observed and predicted values, respectively; and N is the number of observations.

The R^2 index was employed to characterize the degree of collinearity between the predicted and observed values and to describe the proportion of the variance that the prediction model could explain. A higher R^2 value often expresses the lower error of the variance. The parameter d proposed by Willmott (1981) is a normalizing measure index concerning the extent of forecast error and variation of the model. The d value ranges from 0 to 1. The complete agreement between the outcome of predictions and observations results in a value of 1. A higher d value corresponds to a better consistency in terms of the predicted and observed data. The d value index is sensitive to detect the additive and proportional variances between the predicted and observed values, while the R^2 value index is insensitive. Moreover, the R^2 and d indices are equally hypersensitive to extremely observed values (Legates and McCabe Jr, 1999). Another performance index, *RSR* (ranging from 0 to ∞), was employed to standardize the model error estimates in combination with an statistical index of root mean square error and the standard deviation of observations (Moriasi et

al., 2007). A lower *RSR* value indicates a better model simulation accuracy. The model prediction result is regarded as perfect when the *RSR* value is calculated as 0. Meanwhile, the value of root mean square error for a well-described error index is 0.

3 Results

3.1 Descriptive statistics of environmental variables

A statistical description of snowfall and climatic, geographical and topographic variables during the model calibration and validation periods during 1960–2015 is presented in Table 1. Highly similar statistical results for each variable are shown. It is possible that there was no significant difference between the two groups (calibration and validation) of variables (P>0.05) based on the results of variance analysis (ANOVA). This similarity revealed that the validation dataset can be representative of the calibration dataset. The response variable, i.e., snowfall, had higher standard deviation (SD) and coefficient of variation (CV) values, indicating a greater variability in snowfall over the year during both calibration and validation periods. The mean value of temperature was negative, resulting in a negative CV value. The CV value of temperature exhibited the strongest variation among variables. In comparison, relative humidity exhibited a moderate variation and geographical factors (longitude and latitude) showed a lower variation with CV values less than 7.00%. The CV values for other topographic properties ranged from 54.85% to 162.53% in both datasets, indicating that topographic factors had a greater heterogeneity in the study area.

Table 1 Descriptive statistics for snowfall and climatic, geographical and topographic factors for the calibration and validation datasets

Statistic		Snowfall (mm)	T (°C)	RH (%)	WS (m/s)	Lo (°)	La (°)	Ele (m)	Aspect (°)	Slope (°)
Max	C	149.10	7.09	197.17	7.47	94.70	44.97	3545.00	353.66	34.90
	V	133.30	7.05	77.84	7.52	94.70	44.97	3545.00	353.66	34.90
Min	C	0.10	–17.68	33.47	0.36	75.25	39.47	–8.00	8.13	0.75
	V	0.10	–16.50	33.77	0.32	75.25	39.47	–8.00	8.13	0.75
Mean	C	31.75	–0.85	60.52	1.93	84.70	42.74	1216.75	177.25	4.05
	V	29.79	–0.98	59.22	2.08	84.78	42.65	1234.64	172.82	4.08
SD	C	30.47	4.24	13.28	1.11	5.28	1.52	721.19	100.49	6.41
	V	29.02	4.53	11.10	1.33	5.30	1.46	769.66	94.80	6.62
CV (%)	C	95.99	–498.26	21.94	57.50	6.24	3.55	59.27	56.69	158.38
	V	97.39	–460.40	18.75	63.84	6.25	3.43	62.34	54.85	162.53

Note: T, temperature; RH, relative humidity; WS, wind speed; Lo, longitude; La, latitude; Ele, elevation; Max, maximum; Min, minimum; SD, standard deviation; CV, coefficient of variation; C, calibration (n=632); V, validation (n=274). The units of maximum, minimum, mean and standard deviation are consistent with the variables.

3.2 Importance of predictor variables

Table 2 shows significant correlations of snowfall with all variables (P<0.05) excluding wind speed, although most r values have a lower number ($|r|$<0.50). Climatic variables, including relative humidity and temperature, were strongly positively and negatively correlated with snowfall, with r values of 0.64 and –0.53, respectively. A moderate correlation was obtained between latitude and snowfall (r=0.38), whereas the remaining predictor variables were weakly (positively or negatively) correlated with snowfall (r<0.22). The results indicate that the relationships between the predictor variables and snowfall may be difficult to model as an absolutely linear correlation. However, the RF model provides a non-linear estimation method to explore their correlations. The relative importance of the selected predictor variables in the RF model is presented in Figure 2. The relative humidity, temperature and longitude are the three most important predictor variables influencing snowfall variance. The correlation analysis results also showed the highest relevance for relative humidity and temperature. The other predictor variables, including elevation, aspect, latitude and slope, are of secondary importance for snowfall

dynamics. Although the ranking of wind speed was identified as the lowest, its minor importance for snowfall variance should not be ignored.

Table 2 Correlation coefficients between snowfall and predictor variables

	Temperature	RH	Elevation	Longitude	Latitude	Aspect	Slope	Wind speed
P value	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.367
r	–0.53	0.64	0.12	–0.21	0.38	–0.20	0.15	0.03

Note: RH, relative humidity; **, significant correlation at $P<0.05$ level.

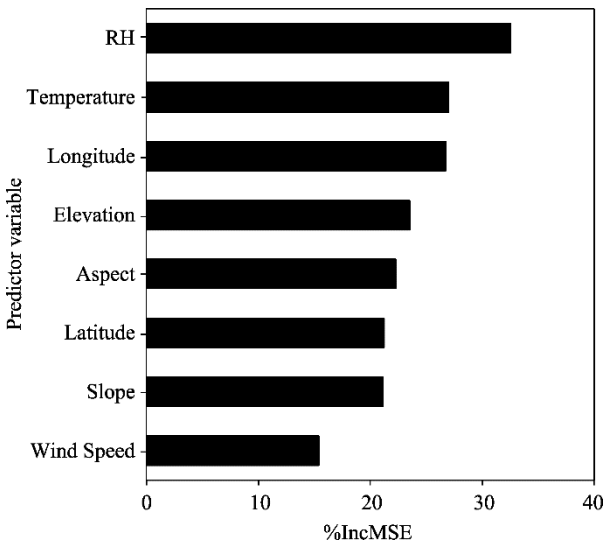

Fig. 2 Importance of predictor variables obtained from the random forest model. RH, relative humidity; IncMSE, mean decrease in accuracy.

3.3 Performance of the MLR model

The model parameters for the eight predictor variables obtained from the MLR model are presented in Table 3. A moderate correlation between snowfall and covariates was derived, with R^2 value of 0.41 (adjusted R^2=0.40), implying that 40% of the total variance in snowfall can be explained by the MLR model. All estimated parameters for the selected predictor variables displayed a high level of significance ($P<0.001$), except for elevation, wind speed and slope, which had a low significance level ($P>0.100$). According to the *lmg* sorting results, relative humidity, temperature and longitude had greater contributions in the regression model, followed by latitude, aspect and elevation, while wind speed and slope contributed the least to the regression model with *lmg* values less than 1.000%.

3.4 Performance of the RF model

The *ntree*, *mtry* and *nodesize* parameters in the calibration period of the RF model were 500, 3 and 5, respectively, according to the selection standard for the specific parameter (see Section 2.3). A larger proportion of the total variance in snowfall was explained in this model, with R^2 values of 0.70 and 0.74, *d* values of 0.90 and 0.93, and *RSR* values of 0.55 and 0.51 during the calibration and validation periods, respectively (Table 4). These results clearly showed the desirable performance and prediction capability of the RF model. The observed and predicted values of snowfall for the validation samples from the RF and MLR models are presented in Figure 3. Generally speaking, the RF model performed much better than the MLR model in predicting snowfall. The predicted results are better when the observed snowfall values are less than 30 mm in both models. However, the overall predicted snowfall amounts were overestimated for low values and underestimated for high values.

Table 3 Parameter estimation results from the multiple linear regression model

Variable	Estimate	Standard error	*t* value	*P* value	*lmg* (%)
Constant	–10.119	56.747	–0.178	0.858	
RH	0.319	0.098	3.248	0.000	8.391
Temperature	–2.104	0.410	–5.121	0.000	7.978
Longitude	–2.561	0.271	–9.439	0.000	7.864
Latitude	5.835	1.210	4.821	0.000	6.149
Aspect	–0.072	0.012	–5.997	0.000	5.152
Elevation	0.002	0.002	0.553	0.580	3.341
Wind speed	–0.694	0.975	–0.712	0.476	0.770
Slope	0.098	0.191	0.512	0.608	0.389

Note: RH, relative humidity.

Table 4 Results of the random forest model

Evaluation indices	Calibration sample	Validation sample
R^2	0.70	0.74
RSR	0.55	0.51
d	0.90	0.93

Note: R^2, coefficient of determination; *RSR*, the ratio of root mean square error to standard deviation of observed dataset; *d*, index of agreement.

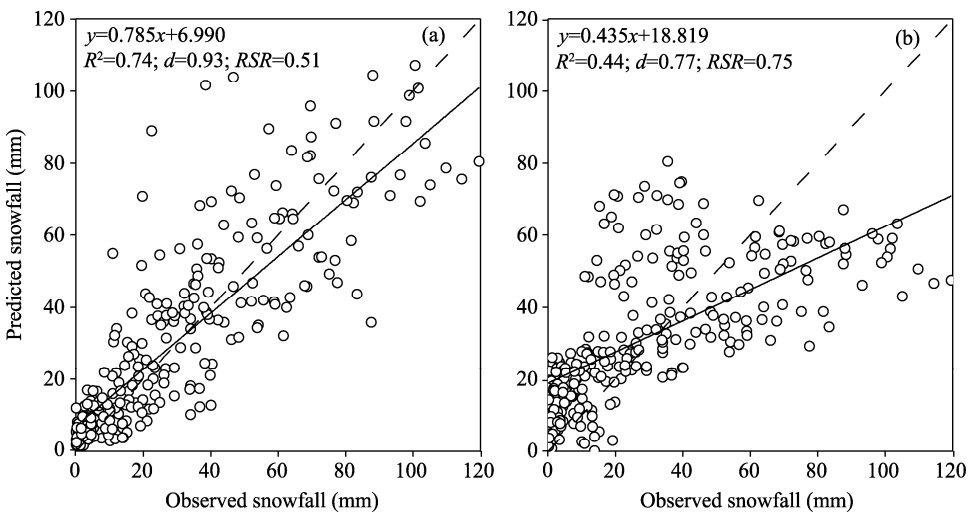

Fig. 3 Correlation of observed and predicted snowfall results obtained from the (a) random forest model and (b) multiple linear regression model using the validation samples

4 Discussion

4.1 Impacts of predictor variables on snowfall

Results from both the RF and MLR models indicate that relative humidity and temperature had particularly important effects on snowfall. Relative humidity and air temperature at low levels are strongly affected by the liquid water amounts and the density of snowflakes at the near surface (Roebber et al., 2003). Karl and Groisman (1993) found that the liquid water equivalent of fresh snow is a function of temperature, relative humidity and other factors. Temperature changes will affect the percentage of precipitation forms and snowmelt regimes, and the amount of snowfall would decrease with a decreasing ratio of snowfall to total precipitation (Kousari et al., 2011; Krasting et al., 2013; Kapnick et al., 2014). Changes in relative humidity may influence the variation in the rate of water vapor formation in the troposphere. O'Gorman and Muller (2010)

and Kousari et al. (2011) found that precipitation evaporation would increase with increasing temperature and decreasing relative humidity. Thus, we propose that changes in temperature and relative humidity may also impact the forms of precipitation. According to Figure 4a, temperature generally exhibited an increasing trend at all the 27 meteorological stations (stations that above the auxiliary line of y=0), with 14 stations showing a statistically significant increasing trend (P<0.05) based on the Mann-Kendall trend test. In contrast, relative humidity showed a decreasing trend in the majority of the 27 meteorological stations (stations that below the auxiliary line of y=0), with 11 stations exhibiting a statistical significance at the P<0.05 level (Fig. 4a). Snowfall changes showed an upward trend at most stations, with only 6 stations showing a statistically significant increasing trend (Fig. 4b). Furthermore, the increase in temperature and decrease in relative humidity were not correlated with a coherent decrease in snowfall for statistically significant stations at elevations below 1500 m. There are two explanations for this observation. First, the significant increase in snowfall in the Tianshan Mountains probably offsets the partial decrease in snowfall arising from the increase in temperature (Guo and Li, 2015). Second, the limited meteorological stations at higher elevation highlight the clear changes in relative humidity and snowfall at lower elevation under the marked increase in temperature.

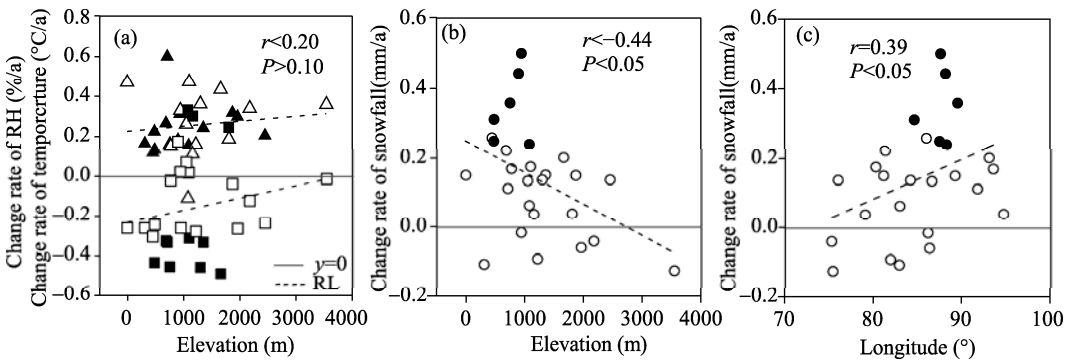

Fig. 4 Scatterplots of (a) change rates of temperature and RH (relative humidity) versus elevation, (b) change rate of snowfall versus elevation, and (c) change rate of snowfall versus longitude. RL, regression line; open boxes (□) indicate change rate of RH at non-statistically significant stations; black boxes (■) indicate change rate of RH at statistically significant stations; Open upward triangles (△) indicate change rate of temperature at non-statistically significant meteorological stations; black upward triangles (▲) indicate change rate of temperature at statistically significant stations; open downward triangles (○) indicate change rate of snowfall at non-statistically significant stations; and black downward triangles (●) indicate change rate of snowfall at statistically significant stations.

Snowfall amount is considered closely related to elevation due to the decrease in temperature with increasing elevation, and temperature can more easily reach the melting point at lower elevation than at higher elevation (Serquet et al., 2011; Rahman et al., 2013). As shown in Figure 4b, there was a significant negative correlation between the change rate of snowfall and elevation. However, the changes in both temperature and relative humidity exhibited a non-significant correlation with elevation. Therefore, the impacts of other potential factors and even microphysical process on the relationship of the change rate in snowfall and elevation should be studied more comprehensively in the future.

Longitude, a geographical factor, was the third most important variable in both the RF and MLR models in our study. The impact of latitude in the RF model was considerably less important, but it ranked the fourth in the MLR model. A significant positive correlation between the change rate of snowfall and longitude is shown in Figure 4c. Latitude can impact the spatial changes in air temperature while longitude may have an influence on wind speed (Kwon and Fu, 2013). Snow may be redistributed by the wind when it falls on the surface (Erickson et al., 2005). Nonetheless, the east-west extension of the Tianshan Mountains is mainly influenced by westerly circulation and wet arctic air masses. Prevailing westerlies can bring moist air masses to the mountains from the Atlantic and Arctic Oceans. However, the high mountains in the west prevent

wet airflow from the westerlies, leading to variations in precipitation with longitude at a local scale (Zhang et al., 2015; Lu et al., 2016). In addition, meteorological stations in the Tianshan Mountains are unevenly distributed in space; hence, longitudinal zonality of snowfall would be more pronounced than latitude under the influence of topography. Similar studies also found a correlation between different forms of precipitation and these two variables (longitude and latitude). Zhang et al. (2015) found that annual precipitation has a distinct longitudinal and latitudinal zonality in the Tianshan Mountains, with an M-shaped distribution of precipitation with longitude and a saddle-shaped distribution of precipitation with latitude. Ji and Chen (2012) revealed that longitude primarily impacts the distribution of precipitation in the middle Tianshan Mountains. Wi et al. (2012) pointed out that in the Colorado River Basin, snowfall shows a significant decrease in all latitude-altitude bands except the highest latitude and altitude.

In this study, aspect was considered as a moderately important factor on snowfall and it ranked the fifth position in the RF model and MLR model. The north and south aspects tend to be dominate in the Tianshan Mountains. The northern and southern slopes are the shade and sunny slopes, respectively. The shade and sunny slopes present different regional climate regimes due to distinct differences in solar radiation, water vapor sources and topographic feature. The shade slope tends to be on the windward side, mainly influenced by westerly airflows with associated precipitation from orographic lifting. The sunny slope is likewise a leeward slope, which is controlled by downward flow with less precipitation compared with the shade slope (Yang et al., 2007; Guo and Li, 2015). Throughout the study period, snowfall and relative humidity on the northern slope were clearly larger than those on the southern slope, while wind speed exhibited the opposite tendency (Fig. 5). Li et al. (2012) also found that winter snowfall on the northern slope of the Tianshan Mountains is significantly greater than that on the southern slope. Moreover, relative humidity is generally higher on the northern slope than on the southern slope, while near-surface temperature lapse rate shows an inverted pattern in this area (Shen et al., 2016). A higher wind speed would decrease the temperature gradient through its influence on air flow. Therefore, the spatial variation in relative humidity, wind speed and near-surface temperature lapse rate influenced by topography and local climate could result in a spatial variation in snowfall between the northern and southern slopes. For both the RF and MLR models, slope and wind speed were the least important variables. However, these two variables are influenced by aspect, which would develop some primary control on snow accumulation and redistribution (Anderton et al., 2004; Anderson et al., 2014).

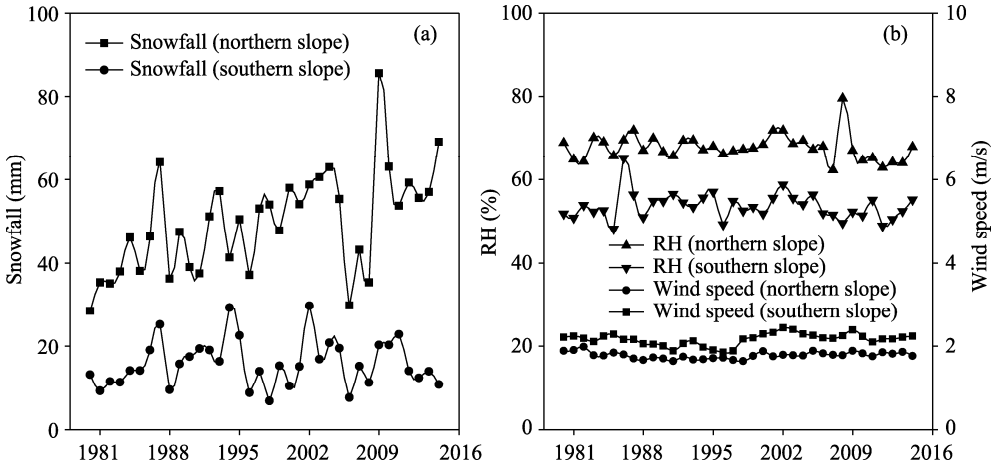

Fig. 5 Variations in (a) snowfall and (b) RH (relative humidity) and wind speed during the period 1980–2015 on the northern and southern slopes in the Tianshan Mountains

4.2 Comparison of the RF and MLR models

Results of the RF and MLR models indicate a distinct difference in model performance during

both the calibration and validation periods. The RF model obtained an obviously higher predictive accuracy than the MLR model, with R^2 values of 0.74 and 0.44, d values of 0.93 and 0.77, and RSR values of 0.51 and 0.75 for the two models, respectively (Fig. 3). The RF model reflected a robust predictive ability, indicating its good applicability in explaining non-linear relationships between the predictor variables and response variable. The MLR model could only partly interpret the variance in snowfall. Notably, the RF model does not need to assume a correlation between the dependent variable and independent variables, and it is less sensitive to some datasets with improper error distributions. Similar comparisons between the RF and MLR models have been proposed by Lopatin et al. (2016) and Zhang et al. (2017) for ecological studies. We found that relative humidity, temperature and longitude were the top three most important contributors to snowfall among all the predictor variables from both the two models. The similar ranking of variable importance in both models showed that climatic factors played a major role in influencing snowfall compared with other influencing variables. It is likely that the RF model could accurately identify the degree of relative importance of variables while it is still being unable to characterize the kinds of effect that these variables may have on the results (Kovdienko et al., 2010). Elevation was the fourth most important variable in the RF model, but it ranked the sixth in the MLR model, which demonstrated that the non-linear relationship was an excellent interpretation of the association between elevation and snowfall. Ning (2013) also found a parabolic-shaped relationship between elevation and mean annual precipitation in the Tianshan Mountains. Temperature will generally vary with the change in elevation. Greater snowfall usually occurs at higher elevation, where the temperature is commonly lower than 0°C.

Varying the out-of-bag samples and rerunning the RF model during the calibration period would generate different rankings of variable importance (Goudarzi, 2016). When the RF algorithm was repeated several times and trained with the same datasets, the first three most important variables were broadly identical at the completion of each model while the rankings of other less important predictor variables changed slightly. In addition, removing lesser important variables, such as slope and wind speed, did not improve the prediction performance of the model. Other studies also found that removing unconnected variables in the RF model had little impact on the results (Palmer et al., 2007). Similarly, in the MLR model, when variables that are insignificantly related or have a minor relative contribution (lmg<1%), such as slope and wind speed, were removed, the adjusted coefficient of determination of the model was similar to the original model, indicating that simplifying the model will not impact the model accuracy.

5 Conclusions

Snowfall prediction is scarce and difficult in the Tianshan Mountains due to regional differences and complex topography. In this study, we utilized different regression approaches to explore the potential driving factors influencing snowfall, i.e., climate, geography and topography. The RF model performed better than the MLR model. It (out-of-bag method) is insensitive to outliers or missing data, and has improved prediction ability for non-linear correlations compared with a similar model (CART). Relative humidity, temperature and longitude were the three most important variables influencing snowfall. Elevation, aspect and latitude were of secondary importance, followed by slope and wind speed. We propose that the impact of climate on snowfall is larger than the topography in general.

Despite the difference in results between the two models, the order of the three most important variables was the same in each model. Furthermore, the significantly important variables listed in this study may be limited, and it may be possible to identify more variables which may influence snowfall. Future studies may focus on identifying more closely related variables to better understand the influences these factors on snowfall in the Tianshan Mountains and to further improve model precision. The RF method could be extended to different regions with superior predictor variables, providing potential insights into regional differences in snowfall amount. In addition, these results will be beneficial to understand hydrological modeling and improve management and prediction of water resources in the mountainous regions.

Acknowledgements

This work was financially supported by the National Key Research and Development Program of China (2017YFB0504201), the National Natural Science Foundation of China (41761014, 41401050) and the Foundation of A Hundred Youth Talents Training Program of Lanzhou Jiaotong University.

References

Anderson B T, McNamara J P, Marshall H P, et al. 2014. Insights into the physical processes controlling correlations between snow distribution and terrain properties. Water Resources Research, 50(6): 4545–4563.

Anderton S P, White S M, Alvera B. 2004. Evaluation of spatial variability in snow water equivalent for a high mountain catchment. Hydrological Processes, 18(3): 435–453.

Antipov E A, Pokryshevskaya E B. 2012. Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2): 1772–1778.

Asaoka Y, Kominami Y. 2012. Spatial snowfall distribution in mountainous areas estimated with a snow model and satellite remote sensing. Hydrological Research Letters, 6(6): 1–6.

Breiman L. 2001. Random forests. Machine Learning, 45(1): 5–32.

Chen Y N, Li Z, Fan Y T, et al. 2014. Research progress on the impact of climate change on water resources in the arid region of Northwest China. Acta Geographica Sinica, 69(9): 1295–1304. (in Chinese)

Chen Y N, Li W H, Deng H J, et al. 2016. Changes in Central Asia's water tower: past, present and future. Scientific Reports, 6(1): 35458.

Chen Y N, Li Z, Fang G H, et al. 2017. Impact of climate change on water resources in the Tianshan Mountians, Central Asia. Acta Geographica Sinica, 72(1): 18–26. (in Chinese)

Clark M P, Andrew G S. 2006. Probabilistic quantitative precipitation estimation in complex terrain. Journal of Hydrometeorology, 7(1): 3–22.

Cutler D R, Edwards T C, Beard K H., et al. 2007. Random forests for classification in ecology. Ecology, 88(11): 2783–2792.

Dai A. 2008. Temperature and pressure dependence of the rain-snow phase transition over land and ocean. Geophysical Research Letters, 35(12): 1–6.

Davis R E, Lowit M B, Knappenberger P C, et al. 1999. A climatology of snowfall-temperature relationships in Canada. Journal of Geophysical Research, 104(D10): 11985–11994.

Erickson T A, Williams M W, Winstral A. 2005. Persistence of topographic controls on the spatial distribution of snow in rugged mountain terrain, Colorado, United States. Water Resources Research, 41(4): 1–17.

Füssel H M, Jol A. 2012. Climate change, impacts and vulnerability in Europe 2012 an indicator-based report. Luxembourg: Publications Office of the European Union.

Genuer R, Poggi J M, Tuleau-Malot C. 2010. Variable selection using random forests. Pattern Recognition Letters, 31(14): 2225–2236.

Goudarzi N. 2016. Free variable selection QSPR study to predict ^{19}F chemical shifts of some fluorinated organic compounds using Random Forest and RBF-PLS methods. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 158: 60–64.

Grömping U. 2006. Relative importance for linear regression in R: the package relaimpo. Journal of Statistical Software, 17(1): 139–147.

Guo L P, Li L H. 2015. Variation of the proportion of precipitation occurring as snow in the Tian Shan Mountains, China. International Journal of Climatology, 35(7): 1379–1393.

Hu R J. 2004. Physical Geography of the Tianshan Mountains in China. Beijing: China Environmental Science Press, 2–4. (in Chinese)

Ikeda K, Rasmussen R, Liu C, et al. 2010. Simulation of seasonal snowfall over Colorado. Atmospheric Research, 97(4): 462–477.

Ji X, Chen Y F. 2012. Characterizing spatial patterns of precipitation based on corrected TRMM $_3B_{43}$ data over the mid Tianshan Mountains of China. Journal of Mountain Science, 9(5): 628–645.

Kapnick S B S, Delworth T L T, Ashfaq M, et al. 2014. Snowfall less sensitive to warming in Karakoram than in Himalayas due to a unique seasonal cycle. Nature Geoscience, 7(11): 834–840.

Karl T R, Groisman P Y. 1993. Recent variations of snow cover and snowfall in North America and their relation to precipitation and temperature variations. Journal of Climate, 6(6): 1327–1344.

Kousari M R, Ekhtesasi M R, Tazeh M, et al. 2011. An investigation of the Iranian climatic changes by considering the

precipitation, temperature, and relative humidity parameters. Theoretical and Applied Climatology, 103(3–4): 321–335.

Kovdienko N A, Polishchuk P G, Muratov E N, et al. 2010. Application of random forest and multiple linear regression techniques to QSPR prediction of an aqueous solubility for military compounds. Molecular Informatics, 29(5): 394–406.

Krasting J P, Broccoli A J, Dixon K W, et al. 2013. Future changes in northern hemisphere snowfall. Journal of Climate, 26(20): 7813–7828.

Kwon T J, Fu L. 2013. Evaluation of alternative criteria for determining the optimal location of RWIS stations. Journal of Modern Transportation, 21(1): 17–27.

Legates D R, McCabe Jr G J. 1999. Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1): 233–241.

Li B F, Chen Y N, Shi X, et al. 2013. Temperature and precipitation changes in different environments in the arid region of northwest China. Theoretical and Applied Climatology, 112(3–4): 589–596.

Li X M, Gao P, Li Q, et al. 2016. Muti-paths impact from climate change on snow cover in Tianshan Mountainous area of China. Climate Change Research, 12(4): 303–312. (in Chinese)

Li X S, Zhang M J, Wang B L, et al. 2012. The change characteristics of winter snowfall, snow concentration degree and concentration period in the Tianshan Mountains. Resources Science, 34(8): 1556–1564. (in Chinese)

Liu Y L, Ren G Y, Yu H M. 2012. Climatology of snow in China. Scientia Geographica Sinica, 32(10): 1176–1185. (in Chinese)

Lopatin J, Dolos K, Hernández H J, et al. 2016. Comparing generalized linear models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. Remote Sensing of Environment, 173(315): 200–210.

Lu H, Wei W S, Liu M Z, et al. 2016. Variations in seasonal snow surface energy exchange during a snowmelt period: an example from the Tianshan Mountains, China. Meteorological Applications, 23(1): 14–25.

Marks D, Winstral A, Reba M, et al. 2013. An evaluation of methods for determining during-storm precipitation phase and the rain/snow transition elevation at the surface in a mountain basin. Advances in Water Resources, 55(3): 98–110.

Mir R A, Jain S K, Saraf A K, et al. 2015. Decline in snowfall in response to temperature in Satluj basin, western Himalaya. Journal of Earth System Science, 124(2): 365–382.

Moriasi D N, Arnold J G, Van Liew M W, et al. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3): 885–900.

Muñoz J, Felicísimo A M. 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science, 15(2): 285–292.

Nair H C, Padmalal D, Joseph A, et al. 2017. Delineation of groundwater potential zones in river basins using geospatial tools—an example from southern western Ghats, Kerala, India. Journal of Geovisualization and Spatial Analysis, 1:5, doi:https://doi.org/10.1007/s41651-017-0003-5.

Ning L K. 2013. Study on the influence of iopography and geomorphology on precipitation over Tianshan Mountains, Central Asia. MSc Thesis. Shihezi: Shihezi University. (in Chinese)

O'Gorman P A, Muller C J. 2010. How closely do changes in surface and column water vapor follow Clausius–Clapeyron scaling in climate change simulations? Environmental Research Letters, 5(5): 025207.

Oliveira S, Oehler F, San-Miguel-Ayanz J, et al. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management, 275(4): 117–129.

Padoan S A, Ribatet M, Sisson S A. 2009. Likelihood-based inference for max-stable processes. Journal of the American Statistical Association, 105(489): 263–277.

Palmer D, O'boyle N, Glen R, et al. 2007. Random forest models to predict aqueous solubility. Journal of Chemical Information and Modeling, 47(1): 150–158.

Perry L, Konrad C. 2006. Relationships between NW flow snowfall and topography in the Southern Appalachians, USA. Climate Research, 32(1): 35–47.

Piazza M, Boé J, Terray L, et al. 2014. Projected 21st century snowfall changes over the French Alps and related uncertainties. Climatic Change, 122(4): 583–594.

Rahman K, Maringanti C, Beniston M, et al. 2013. Streamflow modeling in a highly managed mountainous glacier watershed using SWAT: The upper Rhone River watershed case in Switzerland. Water Resources Management, 27(2): 323–339.

Rasmussen R, Liu C, Ikeda K, et al. 2011. High-resolution coupled climate runoff simulations of seasonal snowfall over Colorado: A process study of current and warmer climate. Journal of Climate, 24(12): 3015–3048.

Roebber P J, Bruening S L, Schultz D M, et al. 2003. Improving snowfall forecasting by diagnosing snow density. Weather and Forecasting, 18(2): 264–287.

Scipiõn D E, Mott R, Lehning M, et al. 2013. Seasonal small-scale spatial variability in alpine snowfall and snow accumulation. Water Resources Research, 49(3): 1446–1457.

Serquet G, Marty C, Dulex J P, et al. 2011. Seasonal trends and temperature dependence of the snowfall/precipitation-day ratio in Switzerland. Geophysical Research Letters, 38(7): 14–18.

Shen Y J, Shen Y, Goetz J, et al. 2016. Spatial-temporal variation of near-surface temperature lapse rates over the Tianshan Mountains, Central Asia. Journal of Geophysical Research: Atmospheres, 121(23): 14006–14017.

Shi Y F, Shen Y P, Kang E S, et al. 2007. Recent and future climate change in northwest China. Climatic Change, 80(3–4): 379–393.

Sorg A, Bolch T, Stoffel M, et al. 2012. Climate change impacts on glaciers and runoff in Tien Shan (Central Asia). Nature Climate Change, 2(10): 725–731.

Strobl C, Boulesteix A L, Kneib T, et al. 2008. Conditional variable importance for random forests. BMC Bioinformatics, 9(1): 307.

Tang Z G, Wang J, Wang X, et al. 2017. Spatiotemporal variation of snow cover in Tianshan Mountains based on MODIS. Remote Sensing Technology and Application, 32(3): 556–563. (in Chinese)

Tinkham W T, Smith A M S, Marshall H P, et al. 2014. Quantifying spatial distribution of snow depth errors from LiDAR using Random Forest. Remote Sensing of Environment, 141(2): 105–115.

Vrotsou K, Fuchs G, Andrienko N, et al. 2017. An interactive approach for exploration of flows through direction-based filtering. Journal of Geovisualization and Spatial Analysis, 1(1–2): 1, doi: https://doi.org/10.1007/s41651-017-0001-7.

Wang L, Liu H L, Bao A M, et al. 2016. Estimating the sensitivity of runoff to climate change in an alpine-valley watershed of Xinjiang, China. Hydrological Sciences Journal, 61(6): 1069–1079.

Wetzel M, Meyers M, Borys R, et al. 2004. Mesoscale snowfall prediction and verification in mountainous terrain. Weather and Forecasting, 19(5): 806–828.

Wi S, Dominguez F, Durcik M, et al. 2012. Climate change projection of snowfall in the Colorado River Basin using dynamical downscaling. Water Resources Research, 48(5): 205–210.

Willmott C J. 1981. On the validation of models. Physical Geography, 2(2): 184–194.

Xu J R, Qiu J Q. 1996. A study on snowfall variation in the Tianshan Mountains during the recent 30 winters. Journal of Glaciology and Geocryology, 18(S1): 123–128. (in Chinese)

Xu L G, Zhu M L, He B, et al. 2014. Analysis of water balance in Poyang Lake Basin and subsequent response to climate change. Journal of Coastal Research, 68: 136–143.

Yang J, Fang G H, Chen Y N, et al. 2017. Climate change in the Tianshan and northern Kunlun Mountains based on GCM simulation ensemble with Bayesian model averaging. Journal of Arid Land, 9(4): 622–634.

Yang Q, Cui C X, Sun C R, et al. 2007. Snow cover variation during 1959–2003 in Tianshan Mountains, China. Advances in Climate Change Research, 3(2): 80–84. (in Chinese)

Yu J Y, Zhang G Q, Yao T D, et al. 2015. Developing daily cloud-free snow composite products from MODIS Terra–Aqua and IMS for the Tibetan Plateau. IEEE Transactions on Geoscience & Remote Sensing, 54(4): 2171–2180.

Zhang F Y, Bai L, Li L H, et al. 2016. Sensitivity of runoff to climatic variability in the northern and southern slopes of the Middle Tianshan Mountains, China. Journal of Arid Land, 8(5): 681–693.

Zhang G, Xie H, Yao T, et al. 2012. Snow cover dynamics of four lake basins over Tibetan Plateau using time series MODIS data (2001–2010). Water Resources Research, 48(10): 10529.

Zhang H, Wu P B, Yin A J, et al. 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. Science of the Total Environment, 592: 704–713.

Zhang X T, Li X M, Gao P, et al. 2017. Separation of precipitation forms based on different methods in Tianshan Mountainous Area, Northwest China. Journal of Glaciology and Geocryology, 39(2): 235–244. (in Chinese)

Zhang Z F, Xi S, Liu N et al. 2015. Snowfall change characteristics in China from 1961 to 2012. Resources Science, 37(9): 1765–1773. (in Chinese)

Zhang Z Y, He H L, Liu L, et al. 2015. Spatial distribution of rainfall simulation and the cause analysis in China's Tianshan Mountains area. Advances in Water Science, 26(4): 500–508. (in Chinese)