Spatial prediction of spring locations in data poor region of Central Himalayas

Rabin Raj Niraula, Subodh Sharma, Bharat K. Pokharel and Uttam Paudel

ABSTRACT

This research explores the methods for understanding groundwater springs distribution and occurrence using Geographic Information System (GIS) and Machine Learning technique in data poor areas of the Central Himalayas. The objectives of this study are to analyse the distribution of natural springs, evaluate three random forest models for its predictability and establish a model for the prediction of occurrence of springs. This study evaluates the primary causal factors for occurrence of springs. The data used in this study consists of 20 parameters based on topography, geology, lithology, hydrology and land use as causal factors, whereas 621 spring location and discharge ($n = 621$) measured during 2014–2016 and 815 non-spring locations (generated by GIS tool) use as supporting evidence to train (80%) and test (20%) the prediction model. Results show that the Bootstrap method is comparatively reliable (92% accuracy) over Boosted tree (64% accuracy) and Decision tree (74% accuracy) methods to classify and predict the occurrence of springs in the watershed. Bootstrap Forest shows the high Prediction rate for True Positive (82% actual spring predicted as a spring) and True Negative (89% actual non-spring predicted as non-spring), and the model seems consistent in both responses. This model was then applied to an independent dataset to predict spring location estimates with 75% accuracy. Therefore, spatial statistical methods prove efficient at predicting spring occurrence in data poor regions.

Key words | bootstrap method, groundwater, prediction model, random forest, springs

Rabin Raj Niraula (corresponding author)
Department of Environmental Science and Engineering,
Kathmandu University,
Dhulikhel, Banepa,
Nepal
E-mail: *robin.niraula@gmail.com*

Subodh Sharma
Kathmandu University,
Nepal

Bharat K. Pokharel
Helvetas Swiss Intercooperation,
Nepal

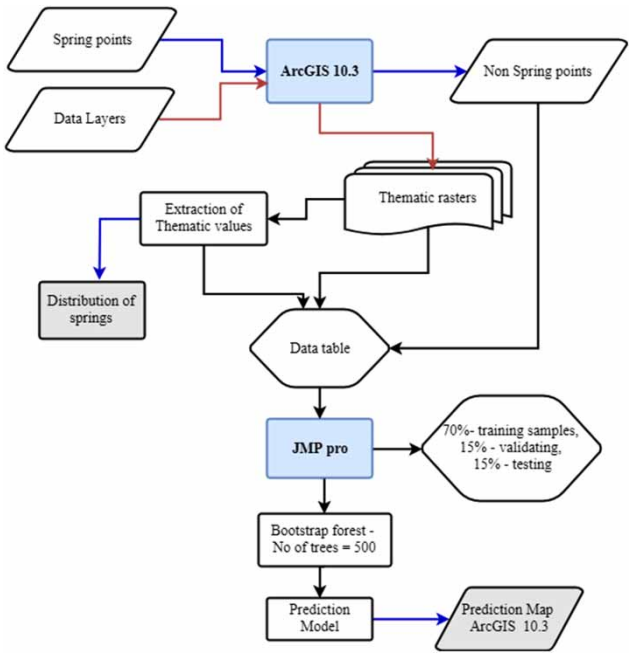
Uttam Paudel
Catholic Relief Service,
Nepal

HIGHLIGHTS

- A novel approach to predict groundwater spring in areas lacking the inventory of groundwater sources.
- High applicability in data poor scenario of Central Himalayas.
- The study identifies elevation as a limiting (redundant) factor to regression problems.
- Results show discharge predictive ability of the model based on the spatial parameter is very poor.
- The model applied to an independent dataset producing promising results.

doi: 10.2166/nh.2020.223

GRAPHICAL ABSTRACT



INTRODUCTION

Springs are the primary source of water in mountainous and hilly areas of the Himalaya. The distribution of the springs and their condition determines the livelihood opportunities of the community, including agriculture, livestock farming as well as provision of clean water for drinking, sanitation and hygiene (Pariyar 2004). Groundwater in the form of mountain springs ensure water security for the majority of the rural population, though springs are mostly overlooked against studies at the basins and sub-basins (Rasul 2014).

Recent problem faced by local communities, mainly drying up of such springs has caused severe problems in such mountain communities (Rasul 2014; Rawat 2014). Water shortages in the central Himalayas occur during the dry periods from March to May sometimes up to mid-June due to low precipitation (Merz *et al.* 2004). Recent climate change studies have come up with results of drying up springs throughout the Himalaya (Gentle & Maraseni 2012; Tiwari *et al.* 2012). A gap of knowledge exists on how the impacts of climate change on recharge mechanism may vary according to aquifers and regions (Meixner *et al.* 2016).

Springs are hydrogeological features defined by geomorphological characteristics (Alfaro & Wallace 1994) fed by groundwater and are largely recharged by rainwater infiltration (Tambe *et al.* 2012). Classification of springs can be deep seated waters and shallow waters into volcanic, fissures, faults and depression, contact and artesian springs (Bryan 1919) but consistent classification is still lacking (Springer & Stevens 2009).

Spatial prediction of groundwater is studied using GIS and Remote Sensing (Ozdemir 2011a); Weight of Evidence and Artificial Neural Networks (Corsini *et al.* 2009); Bivariate statistical model (Moghaddam *et al.* 2013); binary logistic regression method (Ozdemir 2011b) and multicriteria data analysis (Chenini *et al.* 2010). Studies show that groundwater occurrence is controlled by lithology, structures and landforms where GIS and remote sensing proves to be a powerful tool (Solomon & Quiel 2006). A study on groundwater potential modelling considered lineaments, drainage density, topographic wetness index, relief and convergence index as determining factors (Liu *et al.* 2015). Statistical

maps depict the relative probability of occurrence without considering the time factor (Catani *et al.* 2013).

Decision trees can efficiently discover new and unexpected patterns, trends and relationship compared to other spatial techniques. Decision trees are easy to build and interpret and can automatically handle interactions between both continuous and categorical variables. Random Forests (RF) are a combination of tree predictors (Breiman 2001) basically a machine-learning algorithm (Catani *et al.* 2013) for decision-making. Random forests have recently emerged as one of the most commonly applied nonparametric statistical methods in various scientific areas (Shih 2011) and real world applications (Oshiro *et al.* 2012). RFs is widely used in remote sensing and landslide mapping (Brenning 2005; Stumpf & Kerle 2011; Catani *et al.* 2013) due to their good performance. RF belongs to the family of ensemble methods (Genuer *et al.* 2008) and exhibits high accuracy, robustness against over-fitting the training data (Puissant *et al.* 2014) also reduces the noise effect (Breiman 2001).

The objectives of this study are: (i) To compare various 'Random Forest' prediction models and establish a best model to predict spring sources, (ii) To apply and evaluate the predictive model for spring location and discharge based on spatial parameters and (iii) To compare the result of the prediction model in sub-watershed level and evaluate the model by testing in independent dataset.

STUDY AREA AND DATA

The study was conducted in Melamchi watershed in the Central Mid-Hills of Nepal, 40 km north east of the Kathmandu valley (Figure 1). The Melamchi River, a tributary of the Indrawati river in Koshi basin, originates from the high snowy mountain of the Jugal Himal at an elevation of 5,875 m. The length of the river is 41 km and the catchment area of confluence is 324 Km^2. The mean annual flow is 9.7 m^3/s. The climate ranges from sub-tropical in the lower valleys to cool temperate in the upper mountains. The annual average rainfall in the Melamchi basin is about 2,800 mm which is concentrated mostly during four months of the monsoon of mid-June to mid-September.

Jalkanya and Bhimeshwor sub-watershed in Sindhuli district in the Mahabharat range are selected as a testing site due to similarity in topography. The study site and the testing site both lie in the Koshi basin, but varies in topographical, hydrological, and geological condition. This site provides adequate opportunity for testing the method and comparing the results.

Geologically, metamorphic quartzite rocks with soils of colluvial nature dominate the area. The areas is seismically active with frequent earthquake and recent was during 2015 and possesses highly fractured geology. Springs mainly originate from the weathered, jointed, or fractured rock aquifers in the high-grade metamorphosed rocks. The climate of the study area is temperate (mesothermal) with a range of climate from valley to mountain tops in the watershed. Based on the Koppen's classification, the area falls under Cwa or Cwb which demonstrate Monsoon affected Subtropical highland climate with dry winters; coldest month, averaging above 0 °C, all months with average temperatures below 22 °C, and at least four months averaging above 10 °C. At least ten times as much rain in the wettest month of summer as in the driest month of winter (an alternative definition is 70% or more of average annual precipitation received with the warmest six months) (Köppen 1918; Kottek *et al.* 2006). The 12-month rainfall and temperature data of the area based on the nearest climatological station at Nagarkot (Lat: 27.42, Lon: 85.31, elevation: 2163 established: 1971) is studied.

Data collection

This study was conducted during 2014–2016 for data collection and periodic (15 days) discharge data collection for selected 11 springs was carried out during August 2015 to August 2016. The supporting evidence, i.e. the location of springs in the study area was mapped with GPS based field surveys with accuracy of 10 m and discharge measurement was conducted by bucket watch (container/stopwatch) method with average of 3 consecutive measurement records calculating flow using the discharge equation, $Q = V/t$ where Q is the discharge rate calculated based on Volume (V) of discharge collected in time (t). Discharge measurement of springs in mountain topography is difficult (Rawat 2014) and significant creativity and troubleshooting may require on the part of field technicians (Tubman 2013). A total of 621 springs was mapped in the study area as the dependent variable of the study. Similarly, during 2015,

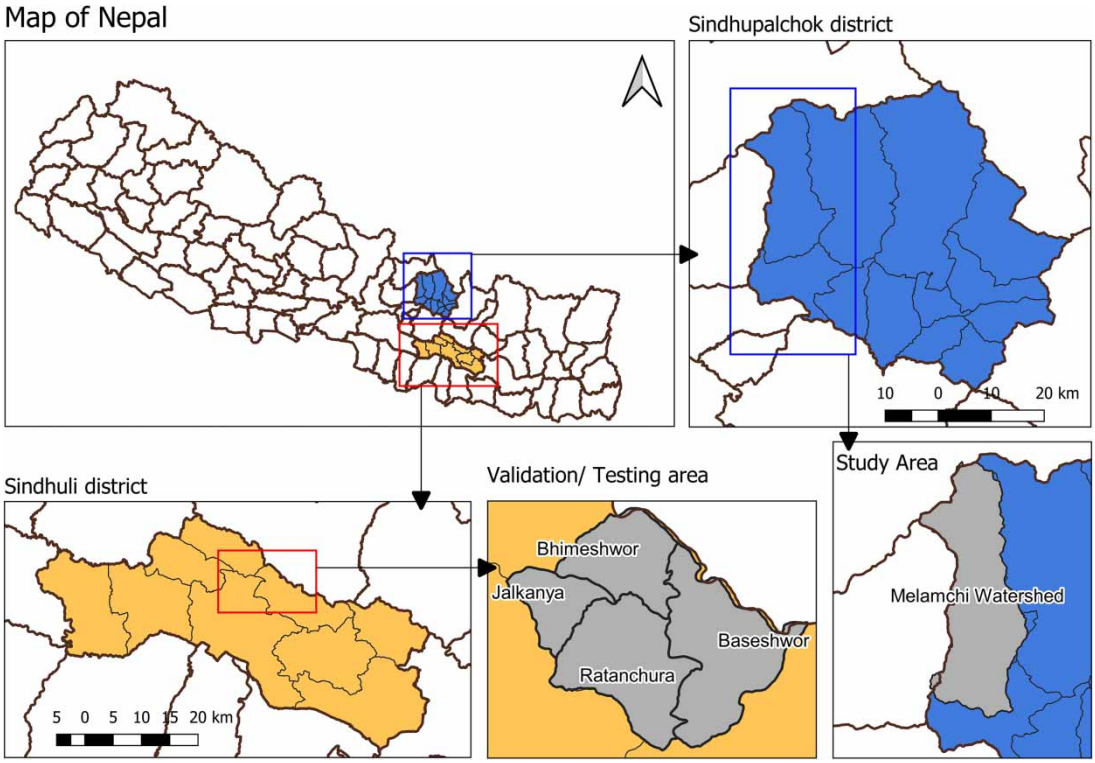


Figure 1 | The study area – Melamchi watershed in Sindhupalchok district and testing area in Sindhuli district.

eighty spring sources were measured at the Sindhuli testing site with the location and discharge of the springs. The testing data were prepared to validate and test the method (data available as Supplementary Material).

The studied springs are located between 1,000 m to 3,000 m of elevation, with discharge ranging from 0.01 litre per second (lps) to 5 lps, with a mean of 0.36 lps as recorded during dry periods (March–May) of the year. The distribution is highly skewed (skewness >1) with high discharge springs being less frequent. High occurrence (67%) of the springs to scatter around 1,000–2,000 m altitude and 37% springs located around 180–270 degrees' aspect (South and South West).

Discharge data of representative 11 springs measured every 15 days for 1 year (Figure 2) clearly suggests that, average discharge of spring measured in *litre per sec* starts to increase from August (mean 0.25 ± sd 0.15) up to October (mean 0.66 ± sd 0.35) and gradually decreases until February (mean 0.22 ± sd 0.12). March onwards the discharge goes as low as drying up in some of the sources which reach the lowest during June (mean 0.08 ± SD 0.07) and slowly starts to rise from July onwards, which is typical for the springs depending on the Monsoon precipitation that is received throughout the country during June to September. The discharge behaviour of these springs suggests that all springs are geologically identical (Bryan 1919) and are recharged in a similar pattern during monsoon as winter precipitation is insignificant.

GIS datasets

The independent variables as causal factors taken for the study are generated from Digital Elevation Model (DEM) with resolution 30 m × 30 m, Land use and Land cover maps from the Department of the survey, Soil Map provided by Soil and Terrain (SOTER) database and Geological Map provided by Department of Mines and Geology (maps available as Supplementary Material). This study uses DEM derived topographic features previously used in spring prediction research (Corsini *et al.* 2009; Chenini *et al.* 2010; Ozdemir 2011a; Moghaddam *et al.* 2013).

Although the DEM-derived parameters represent distinct terrain properties and processes, their interrelationship may lead to multicollinearity. However, for Springs mapping,

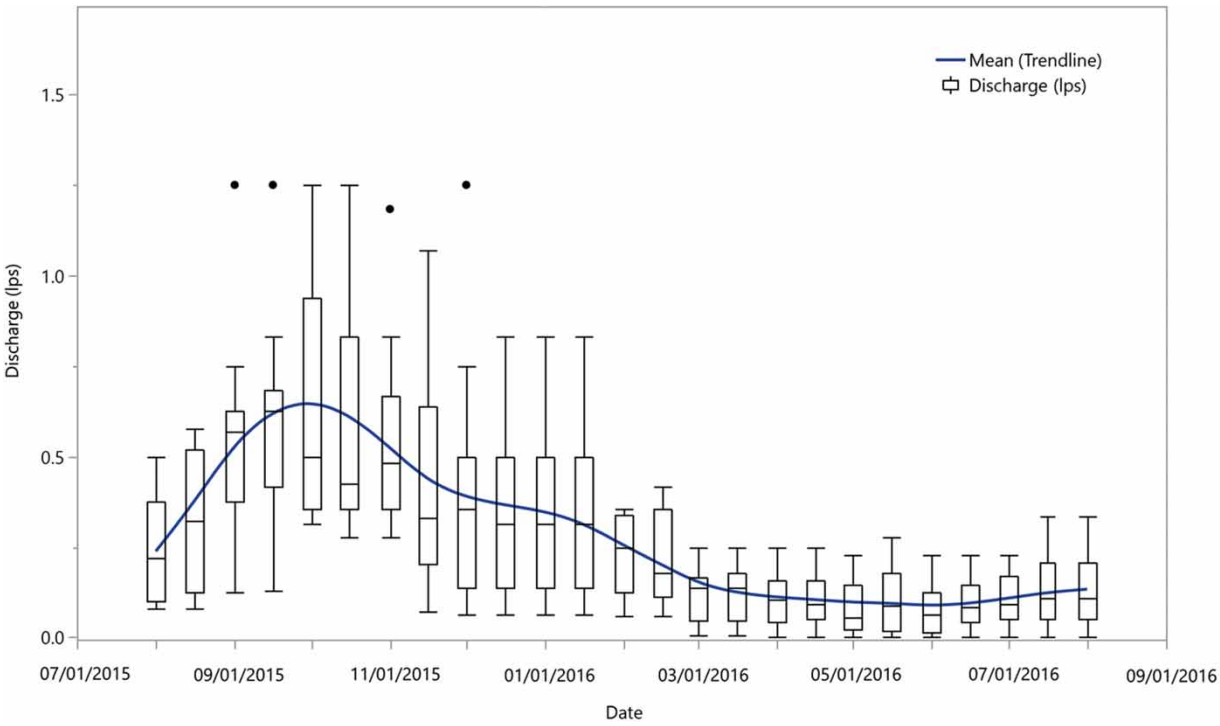


Figure 2 | Spring discharge trend observed during August 2015 to August 2016 in the studied springs. Data available as Supplementary Material.

study (Harrell 2001) suggests that multicollinearity does not influence the predictions from training and testing datasets if both have the same type of collinearities. This applies to this study because the parameters used with the training and testing datasets are mathematical derivatives of the same 30 m DEM. The derivatives are explained in Tables 1 and 2.

Table 1 | Primary topographic attributes derived from a digital elevation model

Data Attribute (Acronym)	Definition/Significance
Elevation (el)	Height above sea level
Aspect (as)	Slope azimuth
Slope (sl)	Inclination of the terrain
Distance to drainage (d2d)	Distance from drainage lines
Drainage density (dd)	Total length of drainage lines per unit area
Total curvature (cr)	Total surface curvature
Plan curvature (plc)	Contour curvature
Profile curvature (pfc)	Slope profile curvature
Distance to ridge (d2r)	Distance from ridge lines
Drop (dr)	Hydrologic slope. Flow, erosion

METHODS

Statistical model

Random Forest (RF) is chosen as the machine learning tool in this study for its superiority in predictive capabilities amongst other present day algorithms (Trigila *et al.* 2015) and it can handle categorical data, unbalanced data as well as data with missing values. Random forest is a widely applied and efficient algorithm based on model aggregation ideas for both classification and regression problems (Breiman 2001). Random Forest is a partitioning method which is good for exploring relationships without having a good prior model, handling large problems producing interpretable results. The predictor variables as well as response variables can be either categorical or continuous (Cutler *et al.* 2012). Random Forest is a supervised learning process which has two steps: Training and Testing. Training involves learning a model using training data samples while the second involves testing the model using remaining data samples to assess the model accuracy. Partitioning is conducted in JMP Pro 12 statistical software where

Table 2 | Secondary topographic attributes derived from a digital elevation model

Topographic Attribute/ (Acronym)	Description
Elevation- relief ratio (hi) (Hypsometric Integral)	The Hypsometric integral (HI) represents the relative proportion of the basin area below a given height or zonal mean. HI = (Hmean- Hmin)/ (Hmax – Hmin) where Hmean = mean elevation, Hmin = minimum elevation and Hmax = maximum elevation
Internal Relief (ir)	Characteristics of terrain roughness
Stream Power Index (spi)	SPI gives the potential of channel erosion and sediment transport process SPI = ln (As × tan B) where As is the specific catchment area
Sediment Transport Capacity Index (stci)	STCI is equivalent to the length-slope factor of the Revised Universal Soil loss Equation $STCI = (m + 1)(A/22.12)^m(\sin B/0.0896)^n$ where A is the upslope contributing area (m^2), B is the local slope gradient (in degrees) and m and n are constants
Terrain Characterization Index (tci)	TCI is related to the spatial variability of soil depth and sediment transportation capacity TCI = Cr lnAs
Topographic Wetness Index (twi)	TWI is related to soil moisture distribution and is useful for groundwater studies TWI = lm (As/ tanB)

groundwater spring data is provided as Response variable 'Y' and topographical, hydrological, geological, soil and land use data are fed as predictor variable 'X'. Random forest can assess the variable importance but cannot show the relationship between the response and independent variables and it should be understood as a predictive tool and a not a descriptive tool.

Training and validation datasets

Classification data used in an RF model for springs mapping should contain information about both springs and non-springs areas. Random points as 'non-springs' were generated in ArcGIS 10.3 software in the study area to provide non-supportive evidence data set against supportive datasets of 'springs point' to avoid over-learning (Corsini *et al.* 2009). Out of 621 known spring points, 3 data were excluded as an outlier and finally dataset included 618 spring points and 815 non-spring points, a total of 1,433 data. The data (57% springs and 43% non-springs) for the study area consist of 1,433 rows each with 20 columns. The data were randomly divided into training (80%) and validation (20%) datasets.

Model evaluation

In statistical classification models, a receiver operating characteristic (ROC) curve evaluates their effectiveness and overall fit (Gorsevski *et al.* 2000). The area under the ROC curve (AUC) characterizes the quality of a prediction model and are used to evaluate the trade-off between true- and false-positive rate of the classification or prediction algorithm (Moghaddam *et al.* 2013). AUC varies from 0.5 (diagonal line) to 1, with higher values indicating a better predictive capability of the model. AUC values less than 0.7 correspond to poor predictive ability, between 0.7 and 0.8 to moderate, between 0.8 and 0.9 to good and >0.90 to excellent (Trigila *et al.* 2010). RF models in this study were evaluated using their predictive accuracy and AUC. A confusion matrix is used to describe the performance of a classification model (classifier or predictor) on a set of test data for which the true values are known which we use in the case of springs and non-springs.

Parameter tuning

Random forest has regression problem in which the range of values response variable can take is determined by the values already available in the training dataset. Unlike linear regression, RF cannot take on value outside the training data. This study identified elevation as a limiting (redundant) factor with regression problem in expanding the prediction model beyond the upper and lower limits of mapped spring sources (1,000 m–3,000 m). To overcome this, for a generalized prediction model, this study excludes elevation as a factor for predicting the occurrence of spring

sources. The results follow this adjustment to improve the prediction model by excluding elevation as a causal factor.

Goodness of Fit

As response variable is categorical, this implies bootstrap partitioning to produce Generalized R-Square (R^2) statistics instead of Mean and standard deviation. The Measure of fit report shows predictors comparison based on R squared statistics, Root Mean Square Error (RMSE) and corresponding Area under curve (AUC) for each model. Generalized RSquare is based on likelihood function L and is scaled to have a maximum value of 1 where perfect predictor has RSquare 1 and 0 for a poor model. Misclassification Rate measures the responses where highest fitted probability differs from the observed responses.

RESULTS AND DISCUSSION

Comparison of prediction models

In this study, Decision Tree, Bootstrap Forest and Boosted Tree methods, are undertaken as powerful predictive models to compare their performances for provided data of 618 springs (excluding 3 outliers) and 815 non-spring points. In all three methods, 80% of 1,433 data were used as training samples and 20% of the same were used as validation samples. As observed, Bootstrap Forest method outperformed other two models with 92% Accuracy based on Area Under Curve (Tables 3 and 4), where Decision tree resulted in 64% accuracy and Boosted tree resulted in 74% accuracy produced as the ability to predict validation data. The null hypothesis of all AUCs produced by 3 models are equal was rejected (Table 5) and the difference between AUCs (Table 6) were also observed to be significant.

Table 4 | Comparison of AUC for springs prediction among all 3 partition model

Predictor	AUC	SE	Lower 95%	Upper 95%
Prob(springs = Yes) Decision Tree	0.64	0.01	0.62	0.67
Prob(springs = Yes) Bootstrap Forest	0.92	0.01	0.91	0.94
Prob(springs = Yes) Boosted Tree	0.74	0.01	0.72	0.76

Table 5 | Hypothesis testing for All AUCs are equal

Test	ChiSquare	DF	Prob > χ^2
All AUCs equal	659.763	2	<.0001*

Based on the Confusion Matrix of the prediction of the data, the Bootstrap Forest method was observed to have high accuracy, precision and predictive ability of the True Positive (TP) as well as True Negative (TN) (Tables 7 and 8) which was comparatively poorly demonstrated by Decision Tree and Boosted Tree. The result indicates that the Bootstrap Forest method is consistent and reliable partitioning method in the prediction of springs location based on the available spatial parameters.

Sub-watershed comparison

Comparison between sub-watersheds is considered as a reliable method to compare the results of the model based on the evaluation of causal factors within a watershed. The data of springs and non-springs was further divided into sub watersheds in this study (Figure 3), as 7 watersheds were selected based on the adequate number of spring data ($N > 30$). The bootstrap forest method could establish prediction model with accuracy ranging from 58% to 100%. In this case, small watershed and insufficient validation data affects the accuracy but this provides comparative

Table 3 | Measures of goodness of Fit

Method	Entropy R^2	Generalized R^2	Mean -Log p	RMSE	Mean abs dev.	Misclassification rate	N	AUC
Decision tree Partition	0.088	0.152	0.624	0.470	0.440	0.38	1,433	0.64
Bootstrap Forest	0.426	0.592	0.392	0.346	0.290	0.13	1,433	0.92
Boosted Tree	0.135	0.226	0.591	0.452	0.427	0.34	1,433	0.74

Table 6 | AUC difference hypothesis testing for three models

Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	χ^2	Prob > ChiSq
Prob(springs = Yes) Decision Tree	Prob(springs = Yes) Bootstrap Forest	− 0.28	0.01	− 0.304	− 0.261	659.76	<.0001*
Prob(springs = Yes) Decision Tree	Prob(springs = Yes) Boosted Tree	− 0.09	0.01	− 0.115	− 0.076	89.645	<.0001*
Prob(springs = Yes) Bootstrap Forest	Prob(springs = Yes) Boosted Tree	0.18	0.01	0.1637	0.2103	246.94	<.0001*

Table 7 | Confusion matrix showing actual versus predicted for all three models

Prediction Method	Decision Tree Predicted		Bootstrap Forest Predicted		Boosted Tree Predicted	
Actual springs	No	Yes	No	Yes	No	Yes
No	0.921	0.079	0.898	0.102	0.904	0.096
Yes	0.773	0.227	0.174	0.826	0.654	0.346

Table 8 | Comparison of performance of selected three models

		Decision Tree	Bootstrap Forest	Boosted Tree
Accuracy	(TP + TN)/Total	0.62	0.87	0.66
Misclassification Rate	(FP + FN)/Total	0.38	0.13	0.34
True Positive Rate	TP/Actual Yes	0.23	0.83	0.35
False Positive Rate	FP/Actual No	0.08	0.10	0.10
Specificity	TN/ Actual No	0.92	0.90	0.90
Precision	TP/Predicted Yes	0.69	0.86	0.73

analysis of similar spatial parameters how they vary in contribution to classify and predict springs and non-springs in the area. Even size of the watershed should be considered as important criteria to establish such prediction models. Comparison of the contribution of spatial parameters in 7 different watershed shows that all parameters have variety of contributions, while Aspect, Distance to drainage, Elevation, Sediment transport capacity index, drop elevation are few having high contribution in most of the watershed (Table 9). It is observed that a prediction misclassification rate of maximum 42% and minimum 0% was produced by the model. The minimum misclassification is resulted when the sub-watershed has least number of validation set (sub-watershed region 26) and highest training accuracy is observed when the sub-watershed has highest number of training set (sub-watershed region 23). This demonstrates that higher number of validation set (here it was 20%) is required to produce reliable prediction model.

Discharge prediction model

Discharge prediction model was tested based on the bootstrap forest method where Response category was based on Discharge data of 621 springs in litre per second categorized in 5 classes to understand the predictive ability of discharge based on provided spatial parameters. This was done to reduce discrete data (discharge) into categorical data. The performance of the model based on 618 springs (3 outliers reduced) is shown in Table 10.

The discharge predictive ability of the model based on the spatial parameter is very poor as the observed misclassification rate of validation set is 62%. This shows that applied spatial parameters are not sufficient to understand and predict the discharge of springs in the hill slope. The subsurface hydrology, below ground geology and characteristic of aquifer is most important to understand the discharge which is not captured due to data unavailability. Data on Aquifer characteristics are not available and complicated which limits the study. So, the model is observed to be weak and not reliable for prediction of discharge.

Spring occurrence prediction model

Random Forest (Bootstrap) method with 20 causal factors generated 500 trees for classification and voting produced

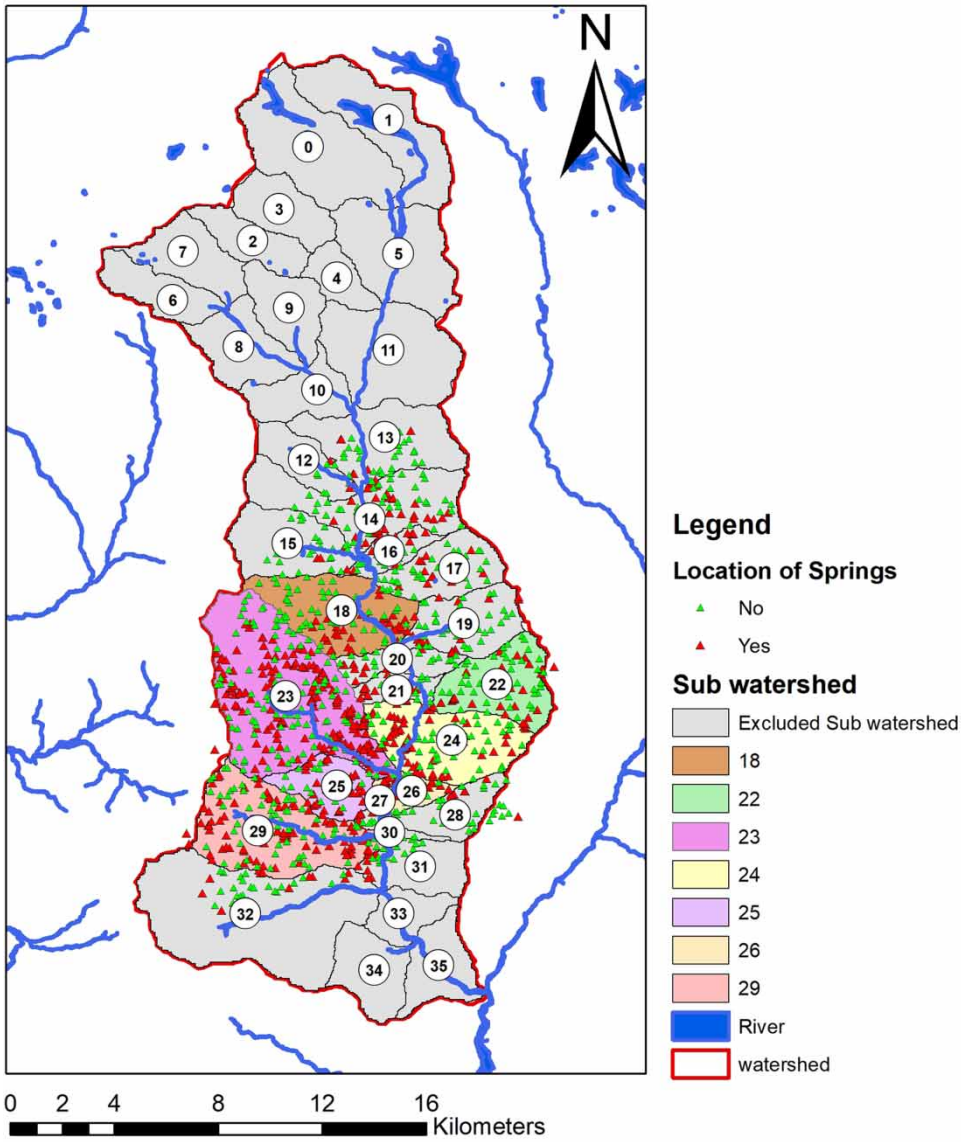


Figure 3 | Map showing sub-watersheds in number and springs locations.

the model with 96% training accuracy and 72% validation accuracy of the data. As this study aims to identify major contributors for the classification and prediction of spring occurrence, the Column contribution statistics shows that Distance to geological boundary with Generalized R^2 49.04 shows highest contribution, while Internal relief (generalized $R^2 = 37.74$), Soil classes (generalized $R^2 = 35.18$), Distance to drainage (generalized $R^2 = 33.84$) and Aspect (generalized $R^2 = 33.36$) are among the highest contributors in the model. Yet, the observation when compared with sub-watershed level contributors shows that not a single parameter can be a major contributor throughout the watershed but there are multiple parameters that interplay.

Additionally, the model showed regression limitation of elevation parameter which resulted in predicting no springs above an altitude of 3,000 m. This is a common error of the method that it cannot train itself beyond training data range. To improve this, the model was re-run with 19 parameters, excluding elevation which resulted in the improved prediction model. Though elevation was excluded, Relief, Hypsometric interval and curvatures are topographic parameter which considers the role of elevation related

Table 9 | Comparison of parameter contribution in selected 7 sub-watersheds

Generalized R Square value

Sub watershed ID		**18**	**22**	**23**	**24**	**25**	**26**	**29**
Area (sq km)		12.38	9.52	27.21	12.82	6.13	2.68	16.90
Number of samples	***N***	*84*	*75*	*274*	*117*	*55*	*54*	*114*
Parameters	**Abv.**							
Aspect	as_30	**8.75**	**4.97**	6.47	2.01	**4.57**	0.00	1.39
Curvature	cr_30	0.00	1.21	2.13	0.00	0.00	**1.84**	**2.63**
Distance to drainage	d2d_30	1.27	**1.98**	**8.92**	1.30	0.20	**3.83**	**2.92**
Distance to ridge	d2r_30	**2.43**	0.30	4.80	1.77	**1.50**	0.00	1.39
Drainage density	dd_30	0.00	0.47	**8.12**	0.00	**1.42**	0.00	2.39
Drop elevation	dr_30	0.00	0.74	7.49	**3.38**	0.00	**6.46**	**3.63**
Elevation	elev_30	0.00	**2.46**	**10.73**	**9.23**	0.00	**4.04**	**2.48**
Geology	geo	0.00	0.00	2.52	1.78	0.00	0.00	1.08
Distance to geological feature	geo_dis	**10.18**	1.93	6.92	0.00	**1.18**	0.00	1.97
Hypsometric interval	hi_30	0.00	1.18	**11.72**	3.12	0.00	0.00	0.38
Internal relief	ir_30	**2.05**	0.66	**7.91**	0.00	1.01	0.00	1.55
Profile curvature	pfc_30	0.00	0.00	2.63	2.42	0.00	0.00	0.59
Plan curvature	plc_30	0.00	1.32	3.30	3.03	0.00	0.00	0.82
Land use	SA_LU	1.51	1.68	4.87	**4.42**	0.00	0.00	1.21
Slope	sl_30	0.00	**2.95**	4.52	0.00	0.00	0.00	2.09
Soil category	soter_ds	1.75	0.00	2.91	0.00	0.00	**4.99**	0.44
Stream power index	spi_30	0.00	0.00	3.51	0.00	0.00	0.00	0.00
Sediment transport capacity index	stci_30	1.40	**2.57**	4.77	**9.20**	0.00	**1.64**	**3.12**
Terrain characterization index	tci_30	**3.03**	0.00	3.69	0.00	0.00	0.00	1.45
Topographic wetness index	twi_30	0.00	0.84	2.79	**5.47**	**1.10**	0.00	1.60

Bold values represent five major contributing parameters.

features. This limitation was not observed in sub-watershed level model as the training data covered whole sub-watershed. This should be considered important in establishing predictive models while preparing data for training samples.

Table 10 | Model performance for discharge prediction

Measure	**Training**	**Validation**	**Definition**
Entropy RSquare	0.3306	0.0285	1-Loglike(model)/ Loglike(0)
Generalized RSquare	0.6514	0.0842	(1-(L(0)/L(model)) ^(2/n))/(1-L(0)^(2/n))
Mean-Log p	0.9670	1.4168	$\sum$ -Log(ρ[j])/n
RMSE	0.6102	0.7294	$\sqrt{\sum}$(y[j]-ρ[j])^2/n
Mean Abs Dev	0.5952	0.7150	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.3732	0.6160	$\sum$ (ρ[j] $\neq$ ρMax)/n
N	493	125	*n*

Above Fit details report of the model in Table 11 shows the classification accuracy for training data is (1–0.0403) × 100% i.e. 96% and the prediction accuracy for validation data is (1–0.2877) × 100%, i.e. 72%. Also, the confusion matrix in Table 12 shows how the cases in the data table were classified and predicted by the current model.

Another important aspect of random forest – Bootstrap method is that it provides estimates of the variable importance shown as column contributions. It shows which variable helps better classify the data for the obtained accuracy. Column contribution sorted in descending order of generalized R square (R2) in Table 13 shows performance

Table 11 | Fit details report of the model

Measure	Training	Validation	Definition
Entropy RSquare	0.4016	0.1274	1-Loglike(model)/ Loglike(0)
Generalized RSquare	0.5674	0.2138	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean-Log p	0.4102	0.5920	$\sum$ -Log(ρ[j])/n
RMSE	0.3441	0.4505	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3287	0.4313	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0403	0.2877	$\sum$ (ρ[j] ≠ ρMax)/n
N	1,141	292	*n*

Table 12 | Confusion matrix for training data of predictive model for spring occurrence

Training			Validation		
Actual	Predicted Count		Actual	Predicted Count	
springs	No	Yes	springs	No	Yes
No	635	6	No	136	35
Yes	40	460	Yes	49	72

Table 13 | Column contribution statistics for comparison of relevance of variables

Term	Number of Splits	(Generalised R Square)	Portion
geo_dis	3111	49.0438362	0.1005
ir_30	2868	37.7458359	0.0774
soter_ds	2661	35.1864185	0.0721
d2d_30	2899	33.8462786	0.0694
as_30	2849	33.3649045	0.0684
dd_30	2534	27.9049329	0.0572
sl_30	2339	26.2276156	0.0538
dr_30	2565	26.0599612	0.0534
hi_30	2436	25.8225324	0.0529
plc_30	2287	24.5406834	0.0503
pfc_30	2139	21.6780314	0.0444
cr_30	2157	21.3123297	0.0437
tci_30	1929	19.9603982	0.0409
twi_30	1938	19.7591879	0.0405
d2r_30	2292	19.3852816	0.0397
spi_30	1901	18.81984	0.0386
stci_30	1857	18.596601	0.0381
SA_LU	1660	16.8928571	0.0346
geo	803	11.6840669	0.0240

of each variable, how it contributes to partition the data based on the model and received accuracy. R^2 (likelihood-ration chi-square) is a statistical test to compare the goodness of fit of two models, one of which (the null model) is the special case of the other (the alternative model). Based on the column contribution, distance to geological features, internal relief, soil, distant to drainage and aspect were the five most influential parameters among those applied. The importance of distance to geological feature is highest ($R^2 = 49.04$) in defining the occurrence of springs.

This result is valid comparing with the findings from similar studies (Ozdemir 2011b) where fault lines were used for prediction of groundwater springs. Additionally, the importance of internal relief, soil type, distance to drainage and aspect is also higher in predicting the occurrence of springs with R^2 (likelihood-ratio chi-square) values equal to 37.74, 35.18, 33.84 and 33.36 respectively. The values are based on the decision tree splits, their performance and partitioning of the data.

The Receivers Operation Curve (Figure 4) explains the training accuracy above 90% is excellent suggesting good separation in the prediction model, whereas the testing accuracy is 77% (between 70 and 80%) and is acceptable. Based on this accuracy assessment, we can accept and apply the prediction model.

Model testing

The model was tested beyond the study area for its potential to replicate in unmeasured areas of similar topography. 80 spring points from Sindhuli were used only as testing samples. The Bootstrap forest method used 1,433 samples from Melamchi (618 springs and 815 non-springs) for training and validation whereas 80 spring points from Sindhuli was exclusively used as testing samples. The bootstrap forest method applied 18 causal factors for the testing with 1,513 samples. The model could accurately predict 75% (Table 14) of the spring points (60 out of 80, see Table 15) even in areas where no training samples was provided suggesting reliability of the model.

CONCLUSIONS

Distribution of natural springs in hill slopes can be affected by many spatial parameters, but it cannot be reflected by any single parameter like elevation or slope aspects, etc. Though springs are formed based on aquifer and geological characteristics, spatial features can reflect the patterns how this occurrence are manifested. In this study, 621 springs were distributed in the hilly slopes of Melamchi watershed,

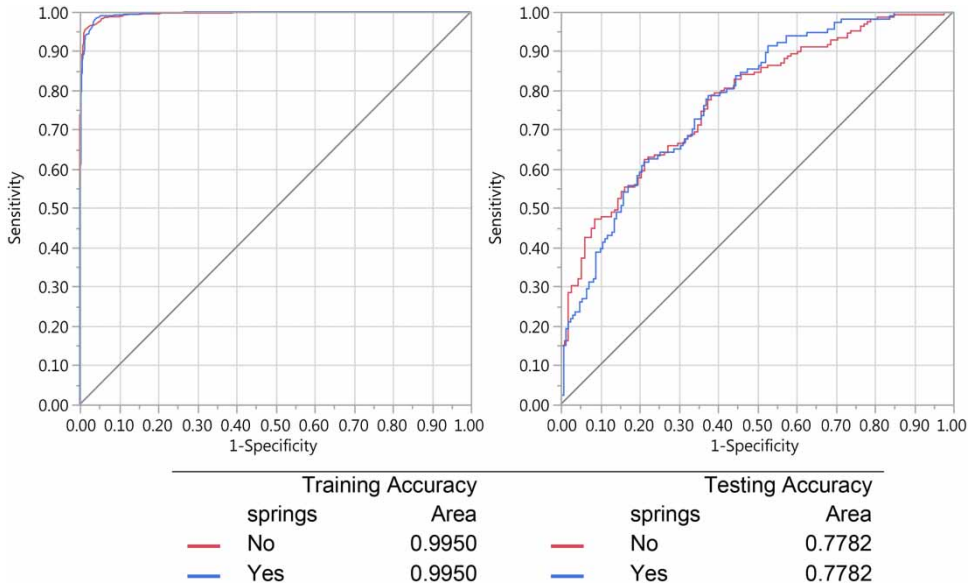


Figure 4 | ROC curve for training and validation data of prediction of occurrence.

Table 14 | Model testing performance statistics

Measure	Training	Validation	Test	Definition
Entropy R^2	0.4573	0.0588	−101.0	1-Loglike(model)/Loglike(0)
Generalized R^2	0.6243	0.1032	−203.4	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean-Log p	0.3720	0.6386	0.6352	$\sum$ -Log(ρ[j])/n
RMSE	0.3207	0.4741	0.4700	$\sqrt{\sum}$(y[j]-ρ[j])^2/n
Mean Abs Dev	0.3021	0.4468	0.4648	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0271	0.3693	0.2500	$\sum$ (ρ[j] ≠ ρMax)/n
N	1,146	287	80	n

Table 15 | Confusion matrix showing testing result

Training			Validation			Test		
Actual	Predicted Count		Actual	Predicted Count		Actual	Predicted Count	
springs	No	Yes	springs	No	Yes	springs	No	Yes
No	641	3	No	125	43	No	0	0
Yes	28	474	Yes	63	56	Yes	20	60

where 20 different thematic layers were studied as causal factors to classify springs and non-spring points, which lacks sub-surface data and rainfall. The springs were more abundant between 1,000 m to 2,000 m elevation (67%) and between 180–270-degree slope aspect (37%). Bootstrap method in Random Forest was observed to have better predictive ability compared to Decision Tree and Boosted Tree method. Bootstrap method as an statistical model can be

applied to prepare the 'spring prediction model' due to its machine learning characteristics, ability to analyse categorical as well as continuous data, high accuracy, robustness against over-fitting the training data (Puissant *et al.* 2014) also reduces the noise effect (Breiman 2001) which was also observed to be effective in this study as compared to Decision Tree and Boosted Tree methods. Prediction of discharge of spring was not reliable as tested in this study. Lack of geological faults data is a major limitation in this study where geological features from Department of Mines, Government of Nepal was applied, yet the data quality is poor in terms of coverage. In this study, elevation was a redundant factor as the recorded location of springs were not beyond 3,000 m altitude, but this limitation was not observed in sub-watershed comparison. Hence in such scenario, the model can predict within the recorded elevation range in the watershed. To overcome this limitation, the model was re-run by excluding elevation. This improved the prediction of model beyond the recorded elevation range of 1,000 m to 3,000 m. Additionally, the same method was tested with 80 spring sources in Sindhuli where the model performed well by accurately predicting 75% of the spring sources. Random forest method is capable of separating provided data into training data and validation data where, validation data is not used for preparing the model but only for the validation of the model which increases the reliability of the results at the cost of reduced testing accuracy. Due to this fact, the model in this study shows 99 percent training accuracy while it shows lower validation accuracy of 72%.

FUNDING

This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada. The views expressed herein do not necessarily represent those of IDRC or its Board of Governors.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Alfaro, C. & Wallace, M. 1994 Origin and classification of springs and historical review with current applications. *Environ. Geol.* **24**, 112–124. https://doi.org/10.1007/BF00767884.

Breiman, L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324.

Brenning, A. 2005 Spatial prediction models for landslide hazards: review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.* **5**, 853–862. https://doi.org/10.5194/nhess-5-853-2005.

Bryan, K. 1919 Classification of springs. *J. Geol.* **27**, 522–561. https://doi.org/10.1086/622677.

Catani, F., Lagomarsino, D., Segoni, S. & Tofani, V. 2013 Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat. Hazards Earth Syst. Sci.* **13**, 2815–2831. https://doi.org/10.5194/nhess-13-2815-2013.

Chenini, I., Mammou, A. B. & May, M. E. 2010 Groundwater recharge zone mapping using GIS-based multi-criteria analysis: a case study in Central Tunisia (Maknassy Basin). *Water Resour. Manag.* **24**, 921–939. https://doi.org/10.1007/s11269-009-9479-1.

Corsini, A., Cervi, F. & Ronchetti, F. 2009 Weight of evidence and artificial neural networks for potential groundwater spring mapping: an application to the Mt. Modino area (Northern Apennines, Italy). *Geomorphology* **111**, 79–87. https://doi.org/10.1016/j.geomorph.2008.03.015.

Cutler, A., Cutler, D. R. & Stevens, J. R. 2012 Random Forests. In: *Ensemble Machine Learning* (C. Zhang & Y. Ma, eds), Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_5

Gentle, P. & Maraseni, T. N. 2012 Climate change, poverty and livelihoods: adaptation practices by rural mountain communities in Nepal. *Environ. Sci. Policy* **21**, 24–34. https://doi.org/10.1016/j.envsci.2012.03.007.

Genuer, R., Poggi, J.-M. & Tuleau, C. 2008 Random forests: some methodological insights. *INRIA* **6729**, 32.

Gorsevski, P. V., Gessler, P. & Foltz, R. B. 2000 Spatial prediction of landslide hazard using discriminant analysis and GIS. In *GIS in the Rockies 2000 Conference and Workshop*.

Harrell, F. E. 2001 Multivariable Modeling Strategies. In: *Regression Modeling Strategies, Springer Series in Statistics*. Springer, New York, pp. 53–85. https://doi.org/10.1007/978-1-4757-3462-1

Köppen, W. 1918 *Klassifikation der Klimate Nach Temperatur, Niederschlag und Jahresablauf (Classification of Climates According to Temperature, Precipitation and Seasonal Cycle). Petermanns Geographische Mitteilungen* **64**, 193–203, 243–248. Available from: http://koeppen-geiger.vu-wien.ac.at/koeppen.htm

Kottek, M., Grieser, J., Beck, C., Rudolf, B. & Rubel, F. 2006 World map of the Köppen-Geiger climate classification updated. *Meteorol. Zeitschrift* **15**, 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Liu, T., Yan, H. & Zhai, L. 2015 Extract relevant features from DEM for groundwater potential mapping. *Int. Arch. Photogramm.*

Remote Sens. Spat. Inf. Sci. - ISPRS Arch. **40**, 113–119. https://doi.org/10.5194/isprsarchives-XL-7-W4-113-2015.

Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., Blasch, K. W., Brookfield, A. E., Castro, C. L., Clark, J. F., Gochis, D. J., Flint, A. L., Neff, K. L., Niraula, R., Rodell, M., Scanlon, B. R., Singha, K. & Walvoord, M. A. 2016 Implications of projected climate change for groundwater recharge in the western United States. *J. Hydrol.* **534**, 124–138. https://doi.org/10.1016/j.jhydrol.2015.12.027.

Merz, J., Nakarmi, G., Shrestha, S., Dahal, B. M., Dongol, B. S., Schaffner, M., Shakya, S., Sharma, S. & Weingartner, R. 2004 Public water sources in rural watersheds of Nepal's Middle Mountains: issues and constraints. *Environ. Manage.* **34**, 26–37. https://doi.org/10.1007/s00267-004-0118-6.

Moghaddam, D. D., Rezaei, M., Pourghasemi, H. R., Pourtaghie, Z. S. & Pradhan, B. 2013 Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan Watershed. *Iran. Arab. J. Geosci* 1–17. https://doi.org/10.1007/s12517-013-1161-5.

Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. 2012 How many trees in a random forest? In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

Ozdemir, A. 2011a GIS-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. *J. Hydrol.* **411**, 290–308. https://doi.org/10.1016/j.jhydrol.2011.10.010.

Ozdemir, A. 2011b Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *J. Hydrol.* **405**, 123–136. https://doi.org/10.1016/j.jhydrol.2011.05.015.

Pariyar, M. P. 2004 In: *Water and Poverty Linkages: Case Studies From Nepal, Pakistan adn Sri Lanka* (I. Hussain & M. Giordano, eds). International Water Management Institute, Colombo, Sri Lanka.

Puissant, A., Rougier, S. & Stumpf, A. 2014 Object-oriented mapping of urban trees using Random Forest classifiers. *Int. J. Appl. Earth Obs. Geoinf.* **26**, 235–245. https://doi.org/10.1016/j.jag.2013.07.002.

Rasul, G. 2014 Food, water, and energy security in South Asia: a nexus perspective from the Hindu Kush Himalayan region. *Environ. Sci. Policy* **39**, 35–48. https://doi.org/10.1016/j.envsci.2014.01.010.

Rawat, P. K. 2014 GIS development to monitor climate change and its geohydrological consequences on non-monsoon crop pattern in Himalaya. *Comput. Geosci.* **70**, 80–95. https://doi.org/10.1016/j.cageo.2014.04.010.

Shih, S. 2011 Random Forests for Classification Trees and Categorical Dependent Variables: an Informal Quick Start R Guide, pp. 1–8. University of California, Berkeley, CA. https://usermanual.wiki/Document/randomForest20guide20in20R.16931

Solomon, S. & Quiel, F. 2006 Groundwater study using remote sensing and geographic information systems (GIS) in the central highlands of Eritrea. *Hydrogeol. J.* **14**, 1029–1041. https://doi.org/10.1007/s10040-005-0477-y.

Springer, A. E. & Stevens, L. E. 2009 Spheres of discharge of springs. *Hydrogeol. J.* **17**, 83–93. https://doi.org/10.1007/s10040-008-0341-y.

Stumpf, A. & Kerle, N. 2011 Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **115**, 2564–2577. https://doi.org/10.1016/j.rse.2011.05.013.

Tambe, S., Kharel, G., Arrawatia, M. L., Kulkarni, H., Mahamuni, K. & Ganeriwala, A. K. 2012 Reviving dying springs: climate change adaptation experiments from the Sikkim Himalaya. *Mt. Res. Dev.* **32**, 62–72. https://doi.org/10.1659/MRD-JOURNAL-D-11-00079.1.

Tiwari, K. R., Balla, M. K., Pokharel, R. K. & Rayamajhi, S. 2012 *Climate Change Impact, Adaptation Practices and Policy in Nepal Himalaya*.

Trigila, A., Iadanza, C. & Spizzichino, D. 2010 Quality assessment of the Italian Landslide Inventory using GIS processing. *Landslides* **7**, 455–470. https://doi.org/10.1007/s10346-010-0213-0.

Trigila, A., Iadanza, C., Esposito, C. & Scarascia-Mugnozza, G. 2015 Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* **249**, 119–136. https://doi.org/10.1016/j.geomorph.2015.06.001.

Tubman, S. C. 2013 *Spring Discharge Monitoring in Low-Resource Settings: A Case Study of Concepcion Chiquirichapa, Guatemala*. Cambridge University Press, Cambridge.

First received 13 August 2020; accepted in revised form 3 December 2020. Available online 29 December 2020