A machine learning based approach for Groundwater mapping

Rashed Uz Zzaman[1], Sara Nowreen[1,*], Irtesam Mahmud Khan[2], Md. Rajibul Islam[2], Nabil Ibtehaz[2], M. Saifur Rahman[2], Anwar Zahid[3], Dilruba Farzana[4], Afroza Sharmin[5], M. Sohel Rahman[2,*]

[1]*Institute of Water and Flood Management ,BUET, Bangladesh,*

[2]*Department of CSE, BUET, ECE Building, West Palashi, Dhaka-1205, Bangladesh,*

[3]*Bangladesh Water Development Board, Dhaka, Bangladesh*

[4]*Department of Public Health Engineering (DPHE), Dhaka, Bangladesh*

[5]*Bangladesh Agricultural Development Corporation, Dhaka, Bangladesh*

[*]Corresponding Author:

Sara Nowreen & M. Sohel Rahman

BUET, Bangladesh

Email: snowreen@iwfm.buet.ac.bd; sohel.kcl@gmail.com

Abstract

In Bangladesh, groundwater is considered to be the main source of both drinking water and irrigation. Suction lift pumps and force mode of operation are the predominant technologies for groundwater abstraction in Bangladesh. For a sustainable usage policy, it is thus important to identify which technology would be more appropriate in which area in Bangladesh. With that aim in mind, this paper proposes a methodology leveraging the power of machine learning (ML) that can potentially learn intricate relationships between the (annual maximum) groundwater level (GWL) and the relevant hydrogeological

factors (HGFs). A number of machine learning algorithms- both classification and regression models- have been trained. Our classification models are trained as a binary classifier to predict the abstraction technology of a particular point. Notably, our best classification model is based on the Random Forest algorithm, which has achieved an accuracy of 91% and an excellent value of 96% for the AuROC (Area Under Receiver Operating Characteristics Curve) indicating a strong discriminant capability thereof. We also identify (elevation derived from) Digital Elevation Model (DEM), Specific Yield and Lithology as the three most important HGFs for GWL in Bangladesh.

On the other hand, to predict the actual (annual maximum) groundwater level, we employ a two-stage approach, where we first employ the above-mentioned classification model to identify the suitable abstraction technology for the point of interest and subsequently predict the actual groundwater level using the appropriate Random Forest regressor and that too with reasonable accuracy (minimum absolute error is less than 1 for suction mode and less than 5 for the force mode). In the sequel, using our predictor models, we prepare groundwater (technology) maps for the whole Bangladesh.

Keywords: groundwater, hydrogeological factors, Machine learning, prediction, suction-mode pump, force-mode pump.

Article Highlights

- A machine learning pipeline for predicting groundwater level (GWL) and abstraction technology is proposed.
- The relationship between the GWL and hydrogeological factors is learned by the proposed models.
- The most influential hydrogeological factors have been identified.
- Ground water (technology) maps for whole Bangladesh have been prepared.

1. Introduction

Groundwater, particularly at the shallow aquifers, is easily accessible, less vulnerable to pollution than surface water (Oke and Fourie 2017), and it is the most essential freshwater resource on the Earth. This underground resource is used mostly for domestic, agricultural, and industrial purposes. Almost 50% of megacities in the world and 80% of irrigation are reliant on groundwater (Bricker et al. 2017). The excessive use of groundwater is resulting in rapid depletion of groundwater level in the aquifer systems thereby creating a threat to the overall sustainability of worldwide water production (Dalin et al. 2018). As groundwater level is an initial indicator to estimate the groundwater quantity (i.e., the net annual recharge mostly for shallow aquifer which is estimated by water table fluctuation method), prediction of groundwater level may aid in the sustainable and effective management of groundwater resources (Hasda et al. 2020; Zhou et al. 2017).

Currently, groundwater is the main source of both drinking water in irrigation in Bangladesh. In the early 1980s, groundwater-fed irrigation became widespread and number of shallow tube-wells (STW) (main source of groundwater supply) increased from 0.1 million to more than 1.5 million (Qureshi et al. 2014). During the past decades, Bangladeshi farmers gradually became more dependent on groundwater as most rivers and canals therein dried up during the dry season (December-May) (Harvey et al. 2006; Shahid 2008). Over-reliance on groundwater resulted in faster groundwater level depletion. Thus, various parts of Bangladesh face water stress, particularly, during the dry season. In this context, predicting groundwater abstraction technology as well as level could prove to be instrumental in sustainable groundwater management. With this brief backdrop, we propose a state-of-the-art machine learning based approach to predict groundwater abstraction technology and groundwater levels in Bangladesh.

There are a few studies in the literature that predict the groundwater level (GWL) principally using time series data and leveraging various soft computing techniques. For example, Husna et al. (2016) predicted the groundwater levels under different time intervals scenarios (i.e., one-week lead, five-week lead, ten-week lead, 15-week lead) using Artificial Neural Networks (ANN). Notably, their study was limited to Dawu Aquifer of Zibo in Eastern China. Very recently, Hasda et al. (2020) predicted the groundwater level in Bangladesh, albeit in a limited setting, focusing only on the Barind tract, situated in northwestern Bangladesh, covering only around 1942 km^2 area. In particular, Hasda et al. (2020) conducted a time series modeling employing a nonlinear autoregressive exogenous model (NARX) that was trained using the Bayesian Regularization (BR) algorithm. They used time series data containing weekly rainfall, temperature, humidity and evaporation during the period 1980–2017 to forecast GWL. Salem et al. (2018) conducted a study, again on a limited scale (i.e., only on a northwestern district of Bangladesh, namely, Rajshahi), where the goal was to analyze the effect of climate change on groundwater-dependent irrigation. As an intermediate output, Salem et al. (2020) predicted the GWL using a Support Vector Machine based model from the projected climactic variables. Salam et al. (2020) explored the relationship between groundwater level and El-Nino Southern Oscillation (ENSO) teleconnection indices during 1981–2017 in the northwestern region of Bangladesh covering 34600 km^2 area. As a sub-aim, they also predicted GWL changes from 2018 to 2025 leveraging ARIMA model. Another recent study by Hoque and Adhikary (2020) made an effort to predict GWL using the weekly GWL and rainfall data leveraging the power of ANN and autoregressive integrated moving average with exogenous variable (ARIMAX) time series models. However, their work is only limited to one of the western districts, namely, Kushtia.

While there are several studies attempting to predict GWL in Bangladesh (as discussed above), most of these studies leveraged historical time series data and or meteorological factors rather than investigating the relationship of GWL with factors influencing groundwater. The groundwater level is assumed to be intricately related to various hydrogeological factors (HGFs). For instance, influence of the thickness and permeability of the upper clay on GWL behavior has been discussed under National Water Management Plan (NWMP) study in Bangladesh (WARPO 2000). On the other hand, in Nowreen et al. (2021), various geospatial-based indicators like lineament density, drainage density, geomorphology, slope, lithology, soil, land use and land cover (LULC), and rainfall have been used to assess groundwater resources in the northwestern part of the country. Findings in recent study (Burgess et al. 2017) indicate the accuracy of the traditional borehole water level measurement as a means to monitor groundwater storage and recharge on the largest fluvio-deltaic aquifer system including the Bengal Delta. However, the intricate relationship between and among these hydrogeological factors and the actual groundwater level has not been hitherto investigated, particularly in the context of predicting GWL. This research gap, particularly in the context of Bangladesh is addressed in this work. Also, most of previous works are limited to a particular area of Bangladesh whereas we here present a country-wide study. Thus, our objective revolves around developing yet another machine learning based model to predict the groundwater level, albeit through taking a detour from the already published works that have mostly focused on time series data and meteorological factors for such predictions and focusing on different hydrogeological factors (i.e., slope, elevation, drainage density, lithology, specific yield etc.) as influential factors for groundwater. This paper makes the following key contributions:

1. We propose a methodology and develop a model leveraging the power of machine learning (ML) techniques to learn the intricate relationships between the GWL and different hydrogeological factors (HGFs).
2. We further identify the most influential factors among the 14 HGFs considered in this study. Our research reveals that (elevation derived from) Digital Elevation Model, specific yield and lithology are the three most important HGFs influencing groundwater in Bangladesh. To the best of our knowledge, this is the first study to identify such important pieces of knowledge thereby extending the knowledgebase of understanding groundwater recharge in Bangladesh.
3. In Bangladesh, two predominant pumping modes/technologies (popularly referred to as groundwater technology), namely, Suction (S) and Force (F) are used for groundwater abstraction. We use our developed ML based (classification) models (using the most influential HGFs) to identify which pumping modes/technology would be appropriate in which area of Bangladesh with promising accuracy.
4. We further develop regression models, again based on the most influential HGFs, to predict the actual values for the GWL through a two-step pipeline for better accuracy.
5. Finally, we produce the groundwater (technology) map (i.e., a map identifying the appropriate abstraction technology) for the whole Bangladesh. In particular, we prepare a 2×2km resolution map for Bangladesh where each grid point is identified using our ML based model as either S (i.e., appropriate for suction mode of operation) or F (i.e., appropriate for force mode of operation). We also prepare a map at the same resolution where the predicted GWL values have been plotted. To the best of our knowledge, this is

the first attempt to prepare a country-wide groundwater technology map as well as GWL map at such a high resolution.

2. Methods

Our methodology evolves around trying to capture this intricate relationship between and among the HGFs and GWL through the power of machine learning. Our goal is to be able to infer for any point, which technology (suction vs. force) would be appropriate. Informatively, suction mode abstraction works when the vertical distance between the centrifugal pump and pumped water level depth is within 7.5 meters; on the other hand, when the pumped water level depth is more than 7.5 meters, we need to apply force mode abstraction. Now, our main idea is as follows. Assume that we have the GWL data labelled as either S (i.e., suction) or F (i.e., force) and HGFs data for a number of points (referred to as representative points hence forth). We can then train a machine learning model as a binary classifier (S vs. F) using the data of these representative points where the HGFs are treated as features.

Figure 1: Overall methodology and the machine learning pipeline.

If there are enough and diversified representative points, we can hope that the model will generalize and successfully learn the intricate relationship we want it to learn. Note that, here first we work on a binary classification problem, where the goal is to infer one of the two classes: S or F. Subsequently, we develop regression models to handle a more difficult problem of inferring the actual ground water levels. To this end, for better accuracy, we employ a two-stage pipeline as follows. We train two regression models, tailored to do well in areas labelled as S and F respectively. When the models are ready, we first use our (binary) classification model to predict the point of interest as either S or F. Based on the prediction, we then employ the appropriate regression model to predict the groundwater level. Figure 1 shows the overall research workflow.

2.1 Study Area and Datasets

The study area of this research is whole Bangladesh and the target is shallow aquifers most of which are unconfined with Holocene deposits with only a few part thereof (~18%) exhibiting semi confined nature for its Pre-Holocene condition (see Figure 2). Thus the principal focus of this study is shallow (unconfined) aquifers where net impact of groundwater stress (either due to less recharge or more discharge) is identifiable by groundwater position (or declination). Lower aquifers are typically located within 10-60 m depth (Ravenscroft et al., 2005) and replenished annually, except in the capital, Dhaka where a significant cone of depression is observed with water table depths of 15 m to 35 m. Outside of Dhaka, seasonal fluctuations are typically up to 8 meters but spatially vary depending on local hydrogeology and groundwater withdrawal (Shamsudduha et al. 2009).

Nationally ~1,400,000 and ~38,000 suction mode and force mode based pumps, respectively withdraw groundwater through irrigation abstraction wells (BADC, 2019). But, suction mode pumps fail when maximum GWL depletes more than 7.5 meters (i.e., the vertical distance between the centrifugal pump and the pump valve is more than 7.5 meters) . That is why the latest available observation points on annual maximum groundwater level is used to facilitate government organizations to prepare policy trajectories for upcoming years. With this in mind, we consider annual maximum groundwater level (GWL) values that occur in April in this study.

Quality assured monitoring observation points of annual maximum groundwater level (GWL) for the year 2018 have been screened from the data collected from different sources, namely, Bangladesh Water Development Board (BWDB) (726 out of 1124 monitoring boreholes), Bangladesh Agriculture Development Corporation (BADC) (2435 out 3164 observation wells) and Department of Public Health and Engineering (DPHE) (414 out of 4831 observation wells) and

Barind Multipurpose Development Authority (BMDA) (25 out of 25 monitoring boreholes). Based on the GWL values, each of these 3600 points has been labelled as either S (i.e., appropriate for suction mode abstraction) or F (i.e., appropriate for force mode abstraction).

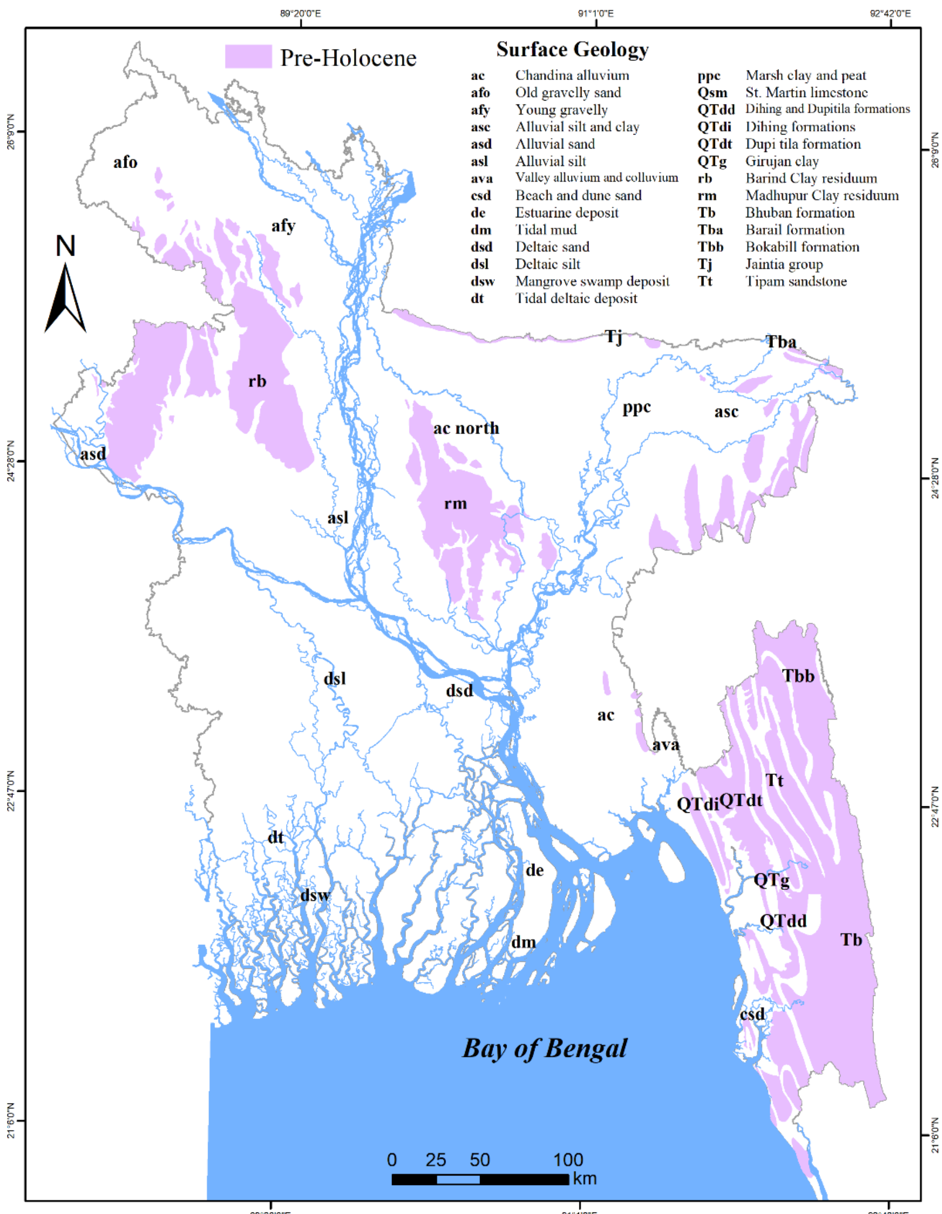

Figure 2: Surface Geology of Bangladesh

2.2 Preparation of hydrogeological factors

Through detailed literature study, a total of fourteen hydrogeological factors-HGFs, namely, Digital Elevation Model (DEM), Curvature, Plan Curvature, Aspect, Slope, Distance from Stream, Lithology, Drainage Density, Stream Power Index (SPI), Sediment Transport Index (STI), Terrain Roughness Index (TRI), Topographic Wetness Index (TWI) and Specific yield (Sy) have been identified that are believed to have intricate relationship with the groundwater level. Specific yield data was collected from BWDB and an interpolated surface was generated using the interpolation technique. Then we proceed as follows.

Freely available lithology map was collected from United State Geological Survey (USGS) (Alam et al. 1990) and subsequently, geo referencing of this map was done in ArcGIS environment. The geo referenced lithology map was digitized to create polygon shapefiles and finally, polygon shapefiles were converted into the raster format in 30m resolution. Digital elevation model (DEM), of 30m resolution, was collected from USGS website ("Earth Resources Observation and Science (EROS) Center" 2021). Notably, elevation has indirect impacts on the groundwater level, as higher elevations have higher slope in our study area which falls into the Hindu Kush Himalayan Region and decreases the infiltration rate (Althuwaynee et al. 2014). Aspect is related to the exposure to sunlight (Lee et al. 2001) and hence has impact over evaporation. Elevation and aspect maps were generated form DEM data using ArcGIS. Slope significantly influences water infiltration and surface runoff (Sarkar et al. 2001). Therefore, we generated the slope map using Equation 1 (below) (Machiwal et al. 2011) in ArcGIS.

$$Slope = 100 \times \sqrt{AX^2 + BY^2} \Big/ _{Pixel\ Size\ (DEM)} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Here, AX (BY) = filtered DEM with x-gradient (y-gradient) filter.

Curvature represents the topography and morphology of the earth surface. It is composed of three aspects, namely, plan, profile and total. Now, the profile and plan curvature mainly impact acceleration and deceleration of flow on the ground surface (Al-Abadi et al. 2016). The plan, profile, and total curvature maps of were generated in ArcGIS using Equation 2.

$$K = \left|\frac{dT}{ds}\right| \quad \text{……………………………………………………….. (2)}$$

Here, T is a unit of tangent vector, ds is a differential of the curves length and | • | denotes the magnitude of the vector.

To generate the drainage density and distance from stream maps, initially the drainage network of was extracted from DEM (30m resolution) by using ArcGIS. The derived drainage network is subsequently used to calculate the drainage density (dd) (Razandi et al. 2015) and distance for stream (d_{ij}) using Equations 3 and 4, respectively.

$$dd = {\sum_{i=1}^{n} Di}/{A} \quad \text{……...…………………………………… (3)}$$

Here, dd is the drainage density $\left({km}/{km^2}\right)$, D_i is the total length of streams and A is the grid area (km^2).

$$d_{ij} = \sqrt{\sum_{k=1}^{n}\left(X_{ik} - Y_{jk}\right)^2} \quad \text{...………………………………… (4)}$$

Here, d_{ij} represents the distance from stream for i, j locations and k means features.

SPI estimates the degree of slope erosion owing to flowing water at a specific location of the basin area. STI measures the sediment transport capacity of overland flow using slope steepness and slope length (Wischmeier and Smith 1978). Following (Sameen et al. 2019), SPI and STI factors were computed using Equations 5 & 6, respectively.

$$SPI = A_s \times \tan\beta \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(5)}$$

$$STI = \left(\frac{A_s}{22.13}\right)^{0.6} \times \left(\frac{\sin\beta}{0.0896}\right)^{1.3} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(6)}$$

Here, A_s is the specific catchment area per unit contour length ($^{m^2}/_{m}$) and β is slope angle (in degrees).

Other influencing factors, namely, TRI and TWI were calculated using Equations 7 & 8, respectively (Sameen et al. 2019).

$$TRI = \sqrt{max^2 - min^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(7)}$$

$$TWI = ln\left(\frac{A_s}{\beta}\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(8)}$$

Here, A_s and β are same as in Equation 6, and max and min are the highest and lowest cell values in the DEM.

2.3 Groundwater abstraction classification using Machine Learning Models

We propose a machine learning (ML) model to classify a point/area (based on the GWL) as characterized by either suction-lift (S) of force-lift (F) abstraction. We have considered the 14 HGFs for this purpose. Figure 1 briefly presents an overview of our machine learning pipeline, which will be further explained in the next subsections.

2.3.1 Selection of Features

It is important to choose features that have strong discriminatory capabilities with respect to classification task at hand as this may have profound effects on the performance thereof. Mostly two types of approaches for feature selection are found in the literature, namely, the filtering and wrapper approach; here we use the former, where a machine learning algorithm (independent of the choice of the actual learning algorithm to do the classification task) is leveraged for feature selection purposes. To this end, we have used the random forest (RF) classifier algorithm for ranking the features. RF algorithm, developed by Breiman (2001), is a nonparametric learning algorithm that generates many classification trees by bootstrap samples thereby attempting to improve the prediction performance. We have used Mean Decrease in Impurity (MDI) as the ranking criterion. MDI refers to the mean total decrease (considering all trees) in node impurities from splitting on the variable. The node impurity is measured by a statistical measure of distribution, namely, Gini index. Higher value of MDI indicates a better feature.

Once all the features are ranked, we try to find the best subset of features for our classification task at hand. We proceed iteratively as follows. We take the most important feature and train (through cross validation) and evaluate our models. Then we extend our feature set by including the second most important feature and so on. For each feature set we have trained our model with several classifiers such as Random Forest, Support Vector Machine (SVM) etc.. Based on different evaluation metrics discussed later on this section, we have found the best subset of features. Notably, we have also used R package ‘leaps’ (Lumley and Lumley 2013) for finding the best subset of features. However, ranking with random forest seems to have served our purpose better.

2.3.2 Training the Classification Model

We have trained our model with different classifiers and have applied *K*-fold cross validation, where the training dataset is first partitioned into *K* equal-sized subsets in order to subsequently train the model with *K*-1 subsets and test it with the remaining subset, repeating this train-test procedure *K* times ensuring that the model is tested against each subset exactly once. In the literature, the popular choice for *K* is 10 which we follow here(Kohavi 1995).

Class imbalance can turn out to be a crucial issue in the context of Machine Learning which is also present in our case as 2413 points are labelled as S (suction-lift) and 1187 points as F (force-lift). Such imbalance in the training data may create a bias in favour of the majority class (i.e., S points). So, in addition to experimenting with the original (imbalanced) instances, we also conducted experiments after applying a popular sampling scheme, called SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al. 2002) on the training dataset thereby oversampling the minority class. However, the results did not change much suggesting that adequate minority class instances were available in the dataset.

2.4 Groundwater level prediction using Regression Models

Following the classification task through the machine learning pipeline presented in the above section, a regression task is also performed to predict the actual value of GWL. Here the goal is to train a machine learning (regression) model using the HGFs as features that can predict the actual ground water level given the HGFs of the respective area/point. For better performance, two separate regression models (i.e., S-Model and F-Model) are trained based on the two abstraction classes (i.e., S and F). So, the goal is to utilize the regression model in a two-step setting: first the classification model is used to classify (using our classification model) whether the area/point under consideration is characterized by suction or force mode abstraction and subsequently leverage the appropriate regression model (i.e., S-Model or F-Model) to predict the GWL value.

The regression pipeline also uses the feature selection step. In particular, our regression models are trained based on the top ranked four features as found through our feature ranking exercise for the classification task. Informatively, any categorical variable is converted to one-hot-encoded vectors following standard procedure. Furthermore, the feature vectors are mean normalized, i.e., normalized to have zero mean with unit variance.

A number of regressor models (Freedman 2005), namely, *K*-Nearest Neighbor (KNN) regressor, Random Forest regressor, Support Vector Regressor (SVR), Adaboost regressor, Neural Network regressor, etc. have been experimented with. We also have performed an ensemble of the models by taking their prediction and fitting a linear regressor with elastic net regularization (Zou and Hastie 2005). Finally, based on the Minimum Absolute Error (MAE) performance measure as well as qualitative inspection of the produced groundwater map, the Random Forest regressor is chosen as the main regressor model. As has been mentioned above, two separate models, namely, S-Model and F-Model, have been trained- one considering the points labelled as S (i.e., having Suction-lift abstraction characteristic) and the other with the points labelled as F (i.e., having Force-lift abstraction characteristic). This makes sense as there are significant characteristic differences between the two sets of points (which was also reflected from significantly worse performance when training was done as a single model considering all points together). The models have been trained following a *K*-fold (K = 10) cross validation scheme. The number of estimators for the Random Forest regressor have been set (through a grid search) to 70 and 100 for the S-Model and F-Model respectively.

2.5 Evaluation Metrics

To evaluate the performance of the classification models, we have used well-established and popular performance metrics from the literature (Altman and Bland 1994; Powers 2020), namely, accuracy, sensitivity, specificity, F1 score, Precision and Matthew's correlation coefficient (MCC). These performance metrics are calculated using the following equations:

$$Accuracy\ (Acc) = \frac{TP + TN}{P + N}$$

$$Sensitivity\ (Sn) = \frac{TP}{TP + FN}$$

$$Specificity\ (Sp) = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Here, *P*, *N*, *TP*, *TN*, *FP*, *FN* respectively denote the number of positives, negatives, true positives, true negatives, false positives and false negatives.

We have also analyzed the ROC-Curve, i.e., the area under receiver operating characteristic curve (ROC-Curve) (Fawcett 2006) and the PR-Curve, i.e., area under precision-recall curve (Davis and Goadrich 2006). These two measures in combination can accurately reflect the performance of a predictor particularly in the context of imbalance in the dataset. To get the ROC-Curve, we need to plot, at various threshold settings, the true positive rate (TPR), i.e., Sensitivity against the false positive rate (FPR), i.e., (1 − Specificity). A ROC-Curve closer to the upper-left corner indicates better performance (Fawcett 2006) and gives a higher (desirable) value for auROC, i.e., the area under the ROC-Curve. To draw the PR-curve we plot, at various threshold settings, the precision against the recall. A PR-curve closer to the upper-right corner indicates better performance of the predictor (Davis and Goadrich 2006) and gives a higher (desirable) value for the auPR, i.e., the area under PR-Curve.

For the regression task, we have used Mean Absolute Error (MAE) as the main performance metric along with the standard deviations. MAE measures the errors between paired observations that express the same phenomenon and is calculated using the following equation:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

In other words, MAE is the mean of the absolute errors, $|e_i| = |y_i - x_i|$, where $|\ y_i|$ and $|\ x_i|$ refer to the prediction and the true value, respectively.

3. Results

3.1 Groundwater abstraction technology classification

Among all the classifiers we have implemented, the best performers turned out to be RF and SVM. We conducted 10-fold cross validation for both of them. The decision threshold was assumed to be 0.5. Table 1 reports the results.

Table 1: Classification performance (on different metrics) of the best models

Classifier	# Features	auROC	auPR	Accuracy	Specificity	Sensitivity	F1 Score	MCC	Precision
Random Forest	4	0.96	0.91	0.91	0.94	0.85	0.86	0.80	0.88
SVM	5	0.90	0.78	0.83	0.92	0.63	0.70	0.59	0.79

For Random Forest and SVM, best number of features are 4 and 5 respectively. Random Forest performs a bit better than SVM. Another important observation is that Random Forest provides a more balanced result as compared to SVM.

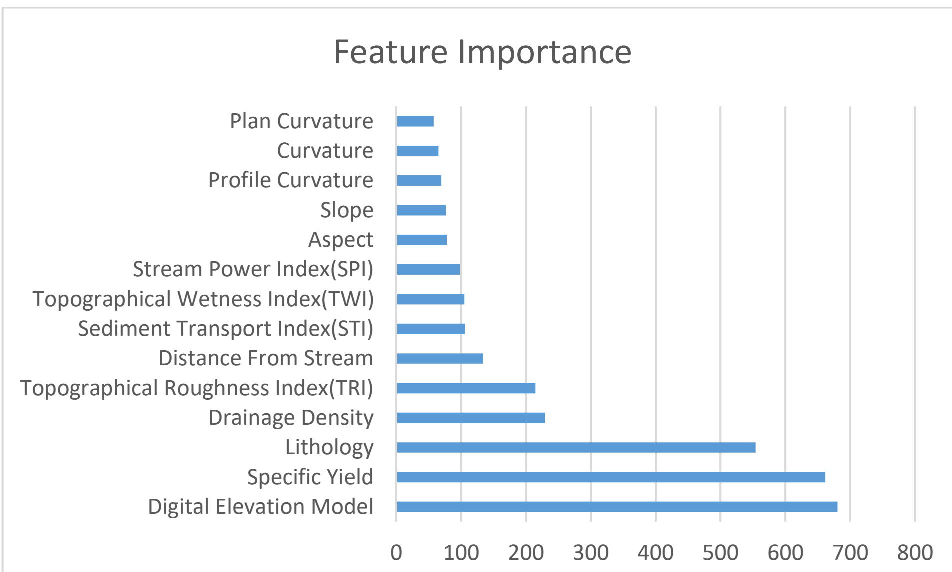

Figure 3: Feature (HGFs) Importance Based on MDI

3.2 Feature Importance

Considering the classification task, we ranked all the features based on Mean Decrease Impurity (MDI) of Random Forest algorithm (Figure 3). We can see that, Digital Elevation Model (DEM) is the most important feature for classification. It is particularly true as elevation affects only renewable (i.e., net annual) recharge part of shallow aquifers of the study under investigation. Also, DEM is closely followed by the Specific Yield (Sy) and Lithology. We further observe that the last few features (e.g., different curvatures) do not have much of an impact on the classification task. Mostly, the top four-five features are sufficient to predict the target.

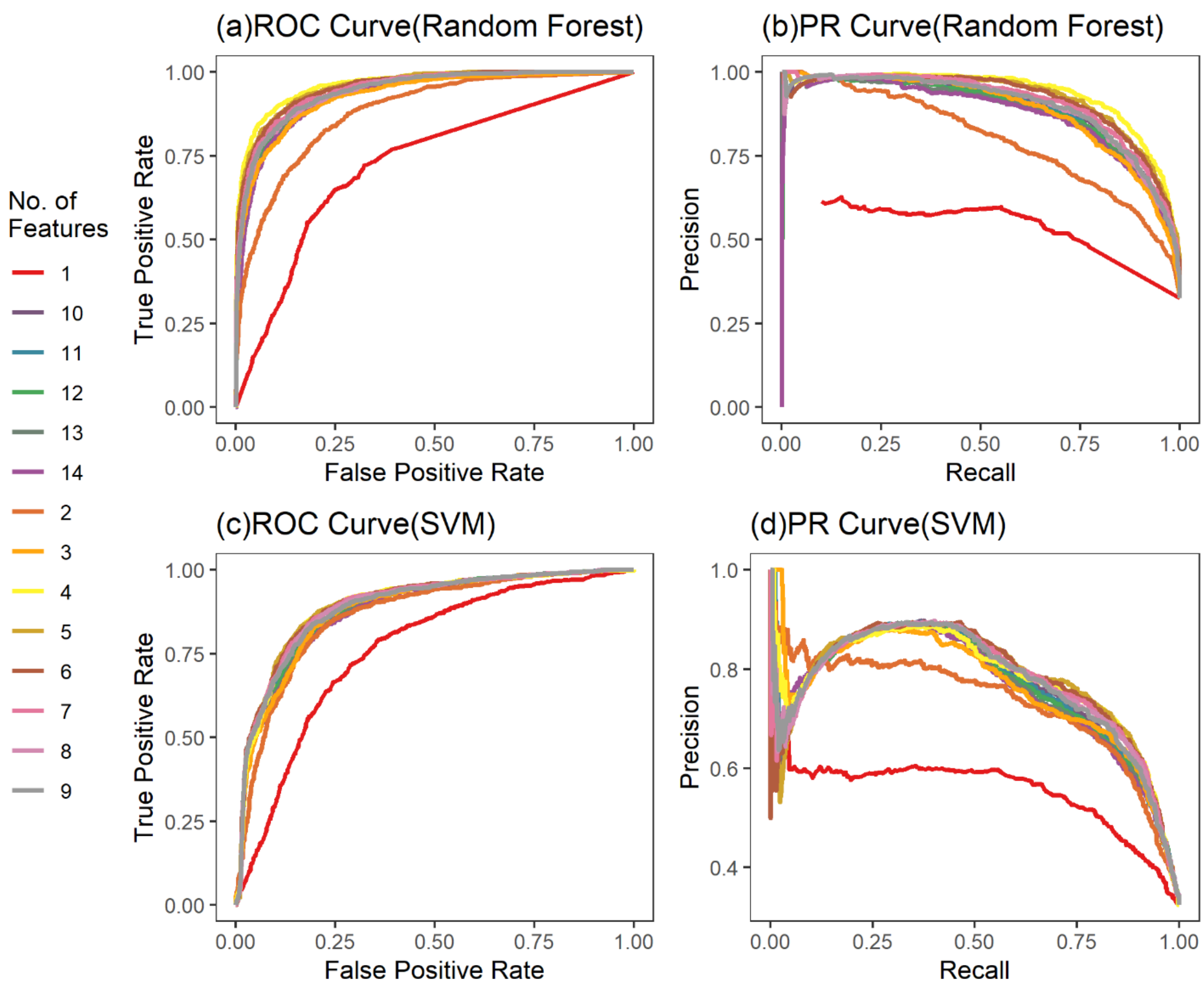

Figure 4: ROC-Curves and PR-Curves for Random Forest and SVM classifiers considering incremental subset of features based on feature ranking. (a) ROC-Curves for Random Forest (b) PR-Curves for Random Forest (c) ROC-Curves for SVM (d) PR-Curves for SVM.

3.3 Impact of Number of Features

To assess the impact of the number of features on the classifier performance, in Figure 4, we plot the ROC-Curves and PR-Curves for both RF and SVM. In each case, 14 different curves are generated as follows: we started with only the most important feature (c.f. Figure 3) and augmented the feature set incrementally by adding the next ranked features one by one. Evidently (c.f. ROC-Curves of Figure 4 (a) and (c)), if we increase the number of features, we notice significant performance improvement at a good rate initially (particularly up to 4 features). Subsequently however, the improvement is not that noticeable i.e., the curves for 4 to 14 features are almost similar. We also analyze the PR-Curve as for imbalanced datasets, ROC-Curve alone is not adequate to assess the impact of selected features; rather PR curve is more relevant in this context (Davis and Goadrich 2006). From the PR-Curves as well (Figure 4 (b) and (d)), we reach the same conclusion.

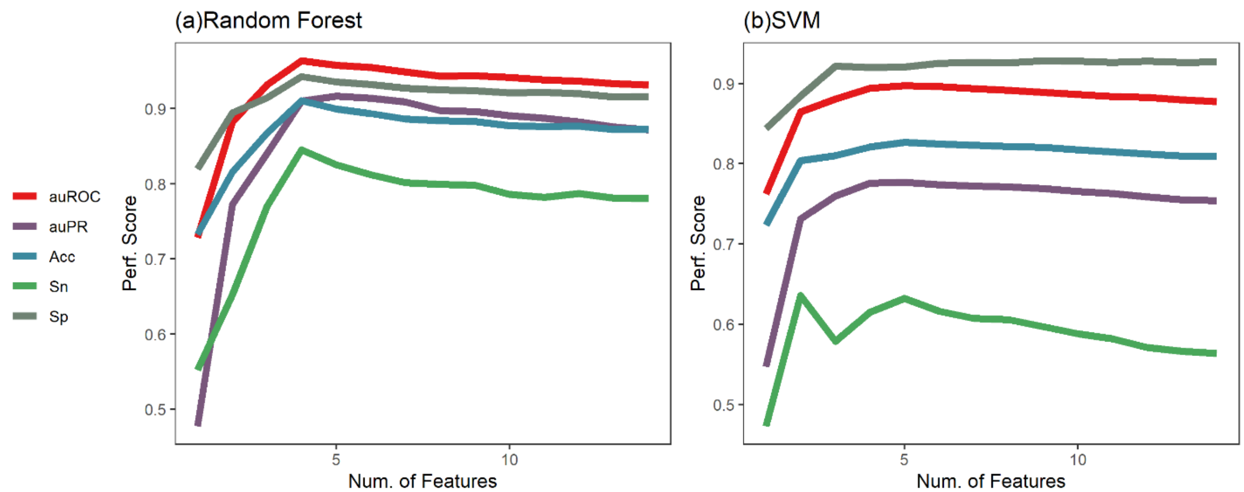

Figure 5. Classifier Performance against different number of features. Performance metrics include area under ROC and PR curves (auROC and auPR), accuracy (Acc), Sensitivity (Sn), Specificity (Sp) and F1 Score (F1). Perf. Score indicates the metric value for a particular performance metrics. (a) Random Forest Classifier performance (b) SVM Classifier performance.

We also have plotted the auROC, auPR, accuracy, sensitivity and specificity values for both Random Forest and SVM models that use varying number of top-ranked features (Figure 5). Like the ROC-Curves and PR-curves, increasing number of features seems to improve all the performance measures albeit up to a certain point, after which, that performance either decreases or gets saturated. Evidently, the best results are achieved using the top 4-5 features.

3.4 Regression results for GWL values

Our Random Forest regressor models, namely, S-Model and F-Model, are trained based on the top ranked four features as found through our feature ranking exercise for the classification task (Figure 3). These features are: digital elevation model (DEM), specific yield (Sy), lithology and drainage density. Lithology, being a categorical variable is converted to one-hot-encoded vectors following standard procedure. The model performances are presented in Table 2.

Table 2: Regression Performance in MAE of our Random Forest regressor models. x $\pm$ y means MAE is x with standard deviation y.

Model	Performance	Comment
S-Model	0.949 $\pm$ 0.05	Applicable for Suction-Lift abstraction points
F-Model	4.296 $\pm$ 0.707	Applicable for Force-Lift abstraction points

3.5 Groundwater (technology) maps for Bangladesh

With the groundwater abstraction (suction-lift vs. force-lift) classification model at hand, we now produce the Ground Water Technology map for whole Bangladesh, which each point/area of the country is identified as having either suction-lift or force-lift abstraction characteristics. To this end, we divided Bangladesh into 2×2km resolution grid as Brammer (2012) reports that, in Bangladesh, there is minimal hydrogeological variation within such a grid. Here, the center point of each grid was used to extract the HGFs values. The extracted HGFs values are particularly chosen from the majority categorical classification of every 2x2km grid using the majority statistical technique known as zonal statistics tool in ArcGIS platform. In fact, we do not need all the HGFs values, we only use the top ranked 4 features (c.f. Figure 3), namely, digital elevation model (DEM), specific yield (Sy), lithology and drainage density (notably, if we decide to run the SVM model we would also need TRI). Subsequently, we have run our Random Forest classification model to produce the Groundwater Technology map for Bangladesh (Figure 6). We have also produced a map where the groundwater level values are presented. To this end, we have used our Random Forest Regressor models (i.e., S-Model and F-Model) in combination with our Random Forest Classification model as follows. As has been presented in Figure 6, our Random Forest Classification model has predicted each 2×2km resolution point as either S (Suction-mode) or F (Force-mode). If a 2×2km resolution point is labelled as S, we predict the value of GWL using the S-Model and otherwise (i.e., if it is labelled as F) we use the F-Model. Figure 7 presents the resulting map (GW-Map).

4. Discussions

High resolution quality data is a prerequisite for proper planning and designing in any sector. To this end, hydrological data collected by BWDB is definitely useful for sustainable water management in Bangladesh. However, the pertinent question in this regard is whether the hydrological data collection network density for the area under consideration satisfies the standard resolution. For Bangladesh, this is certainly lacking: resolution of the groundwater observation well network in the country is about $77km^2$ against the standard resolution of $5\text{-}20/km^2$ (Hossain, M and Zahid 2014). This is inadequate for sustainable groundwater planning and management at the village, or even union (lowest administrative unit) level. High resolution data prediction using machine learning based approaches can fill this gap in this regard.

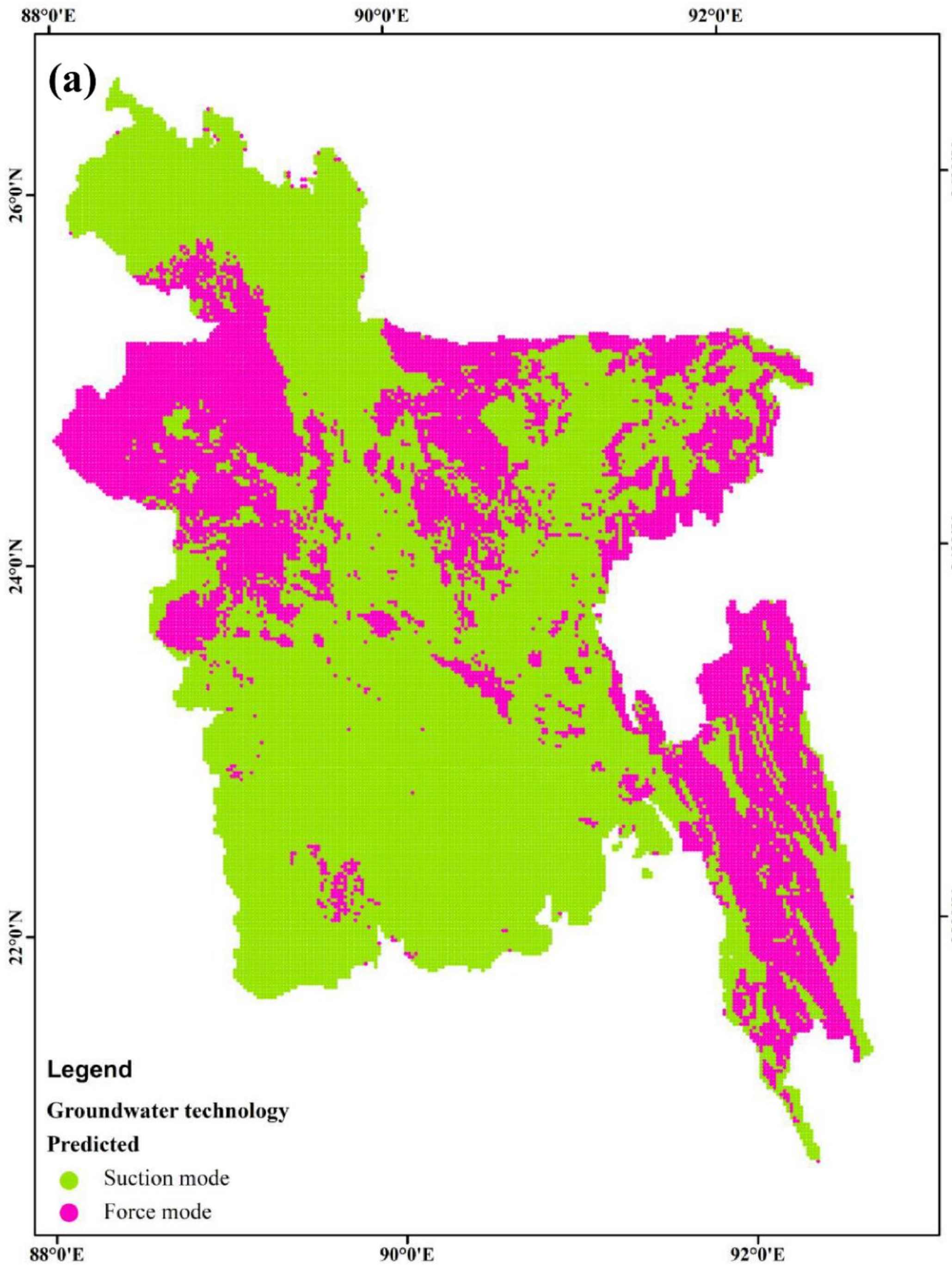

Figure 6: Groundwater (technology) map for Bangladesh. Here suction-mode and force-mode abstraction characteristics have been predicted using our Random Forest classification model in 2×2km resolution grid.

427

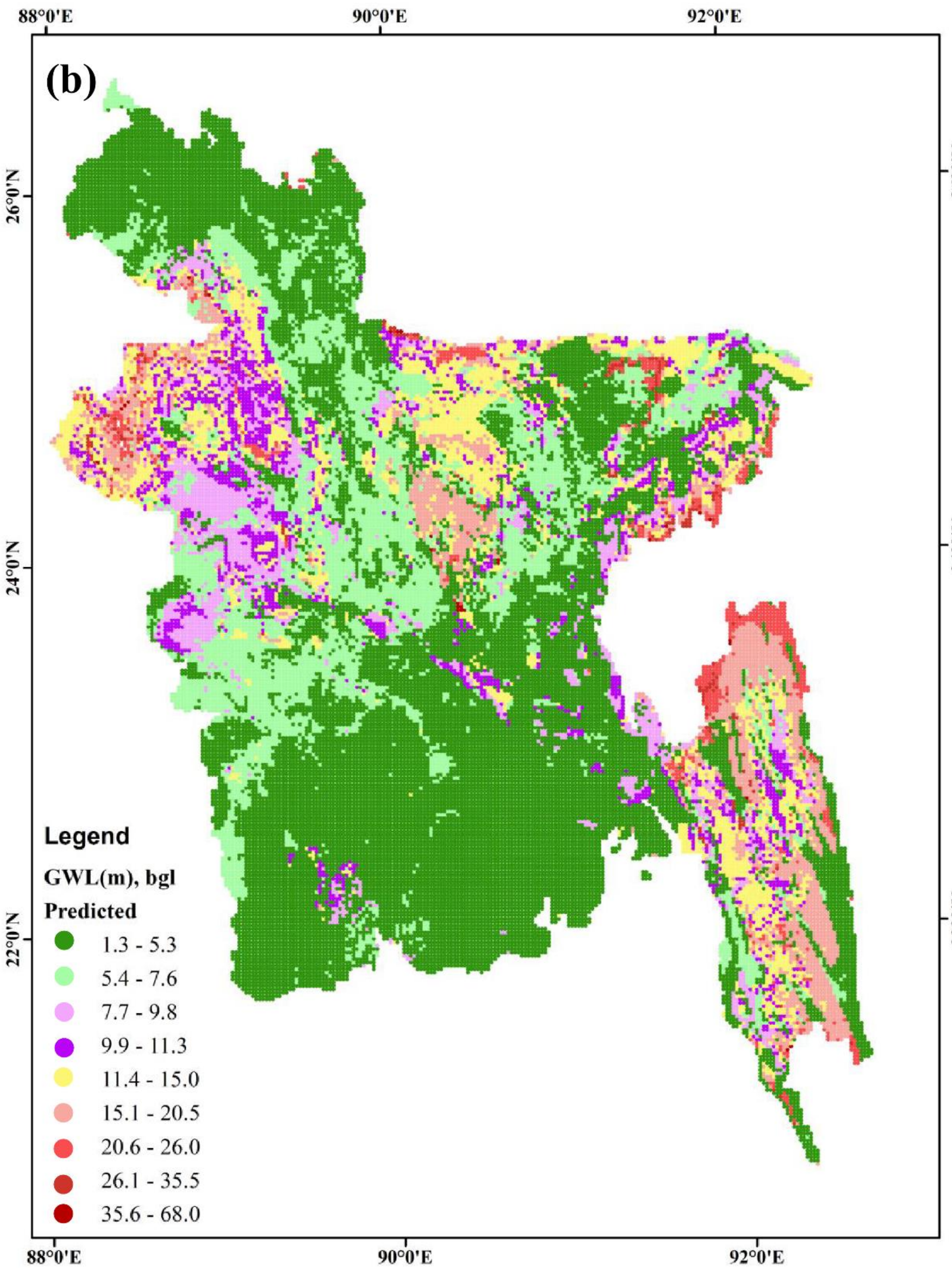

Figure 7: Groundwater map (GW-Map) for Bangladesh. Here annual maximum GWL values that occur in April have been predicted using our Random Forest classification and regressor models in 2×2km resolution grid.

We have leveraged the power of machine learning (ML) models that can potentially learn the intricate relationship between the ground water level (GWL) and the relevant hydrogeological factors (HGFs). A number of studies in the literature have investigated (using different methodologies and approaches) and identified important influencing factors for groundwater. In our research, elevation (DEM) and specific yield (Sy) have been found to be the most influential factors, which are closely followed by lithology. Our findings are in line with that of Arabameri et al. (Arabameri et al. 2020) as they also found lithology and elevation as highly influential. Also, it is well-perceived that in the flood plain, the lithological formation has a big impact rather than other indicators. On the other hand, hilly area indicators are mostly influenced by elevation and slope. Notably, various studies (e.g., Abdollahi et al. 2019; Miraki et al. 2019; Nguyen et al. 2020)) found groundwater aquifer, land use, and TWI as influential factors. Also, LULC, lithology, and elevation were found to be more influential Arabameri et al. (Arabameri et al. 2020). Since 80% of Bangladesh (i.e., our study area) constitute flat area, slope and slope related factors (i.e, TWI, TRI, SPI, and STI) have been found to be less influential. Notably, our selected HGFs already represent the main parts of LULC and hence it is not considered separately in our study.

Using our classification model, we have produced a groundwater technology map for Bangladesh where, in 2×2km resolution, we have predicted each point as having either suction-mode of force-mode abstraction characteristic. Deeper groundwater levels are found to be significantly influenced by impermeable lithology characters, i.e., Barind Clay residuum, Madhupur Tract and hard/rocky layers of the hilly eastern region. In fact, these are the locations where suction mode pumping is usually failing during the dry months (March-April). So extra precaution is always needed before allowing further irrigation expansion with force modes in such areas. From hydrogeological point of view, these sites demand some counteractive actions to prevent recharge loss and avoid groundwater overexploitation (Nowreen et al. 2021). Our produced map therefore could be instrumental in forming and enforcing a sustainable policy in this regard.

Additionally, we have used regression models to predict groundwater levels for Bangladesh. In Figure 7, we present the GWL spatial distribution map of Bangladesh. Because of the lack of infiltration of rainwater, the groundwater level is deeper in a few specific regions (e.g., Barind and Dupi Tial formation) in Bangladesh. This can be attributed to the low transmissivity (in the range of 500–2,000 m^2/day) at the Dupi Tila aquifer system (EPC/MMP 1991) as opposed to the much higher transmissivity (in the range of 3,000– 5,000 m^2/day) at similar depths in the flood-plain Holocene aquifer. GWL is more than 60m BGL (below ground level) in Barind and Dupi Tila formation, which has been captured by our regression model as is evident from some red dots in and around that region. On the other hand, the coastal areas are formed by the recent deltaic formation and groundwater levels in these areas are found to be near from the surface between 0 to 4m. This has also been captured well in our model (Figure 7).

In general, hydrologists commonly predict GWL in unsampled sites by interpolation methods available in the GIS platform, most widely the Ordinary Kriging (OK) method. Now, prediction error variance for the OK method becomes larger where local variability is greatly varied with space, in particular in complex hydrogeological environment (Yamamoto 2000). Figure 8 presents a map (OK-GW-Map) illustrating the spatial distribution of groundwater level all over the country employing Ordinary Kriging using the same training samples used in our ML based approach. Evidently, the OK-GW-Map failed to capture such complex phenomenon in areas like Dhaka (Dupi Tila formation), Barind and Chittagong Hill Tracts (CHT) of Bangladesh which is well-captured in the GW-Map as the machine learning models attempted to learn the intricate relationship between the HGFs and GWL thereby capturing the complex hydrogeological environment in those areas.

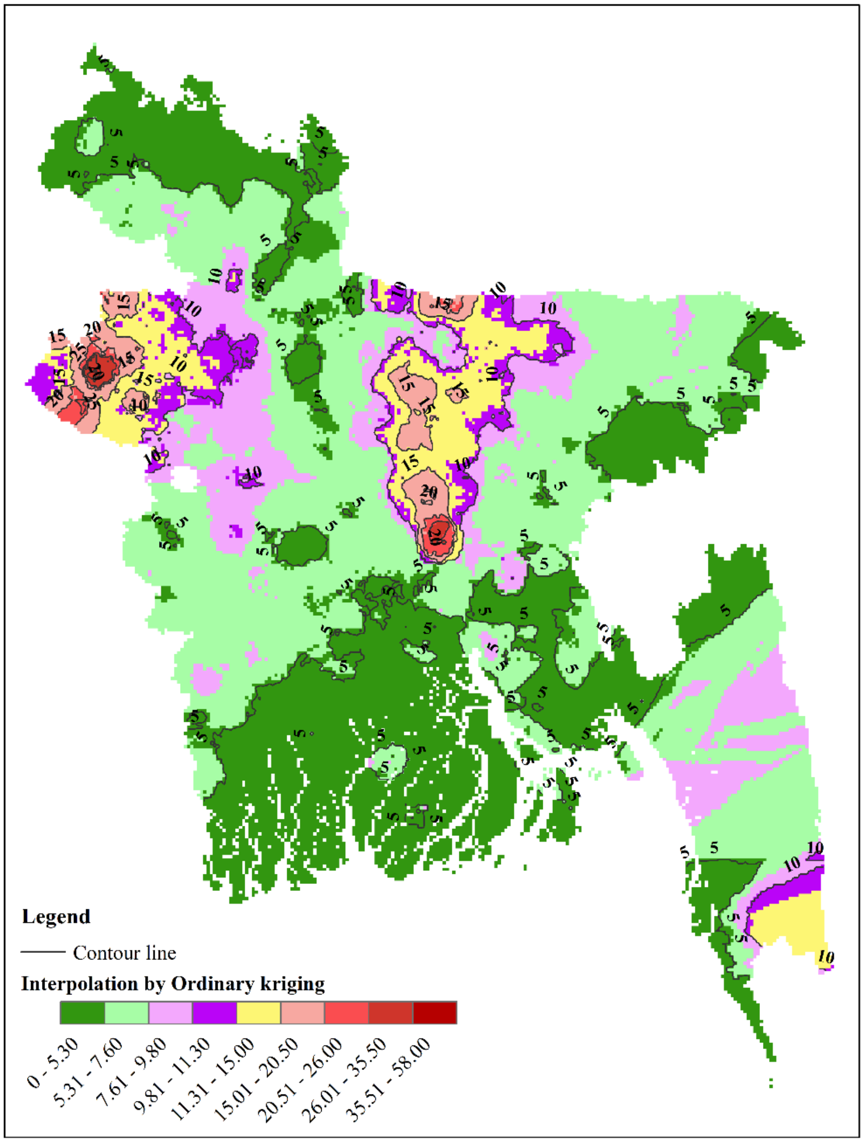

Figure 8: Prediction of annual maximum GWL values that occur in April for Bangladesh by ordinary kriging approach using the same training samples in 2×2km resolution grid.

While we have already discussed the peculiarity of Dhaka (Dupi Tila formation) and Barind area, a brief discussion is also in order including the CHT area. As is evident from the GW-Map, our machine learning models seem to have predicted deeper GWL depths for Dhaka, Barind and Chittagong Hill Tracts (CHT) under the same set of constraints when compared to the OK method and this phenomenon (i.e., deeper GWL depths) is most likely accurate as these three particular regions are well known for their high vulnerability to water crises. Dhaka and Barind’s GWL has deepened owing to high groundwater abstraction for urban water supplies and irrigation, respectively. On the other hand, the dominant shale/clay materials of CHT is what limits the recharge in the subsurface geologic formation and causes higher contour with the groundwater depth. Since there are no samples in the training set with GWL values in the CHT area (due to the absence of observation wells therein), the OK method failed miserably to predict the GWL in those areas (as is evident from the OK-GW-Map), whereas the GW-Map reveals that our models seem to have adequately captured the complex hydrogeological phenomenon in those area.

Our study has inherited the limitations of any machine learning model particularly in the context of the available dataset. The tools and algorithms based on AI and ML are never 100% accurate and the accuracy usually largely depends on the data- both in terms of quality and quantity. Therefore, the output of these approaches must be interpreted wisely and cautiously, specifically where assessment of the dynamic nature of GWL and the spatio-temporal changes are to be taken into account.

5. Conclusion

The ratio of suction mode and force mode abstraction has been significantly changing since the green revolution started in Bangladesh during the eighties. Yet, government organizations like DPHE still suggests pumping technology based on the lowest water table declination forecasting study conducted in 1990 (DPHE 2008), which seems outdated. Therefore, how much withdrawal can really be sustainable with which mode of pumping technology has now become a crucial question for management purposes in many parts of the country. Other developing agrarian countries like India, Vietnam, etc., are no exception. On the other hand, groundwater data collection network density, i.e., standard resolution in Bangladesh is not adequate to forecast groundwater status at the village, or even union level. To this end, our machine learning models could be instrumental in answering such and other relevant questions. Furthermore, we believe that the outcome of this study can aid the policymakers in formulating policies for ensuring sustainable groundwater management. The output of this study will also be instrumental to the policy/decision makers to mark suitable locations for drilling production wells, which, in the sequel, will help farmers reduce the extra unnecessary well construction costs.

DECLARATIONS

Funding

This work is part of the project titled ‘Development of IoT enabled data logger to monitor groundwater and analysis of the collected data’ under the innovation fund of ICT Division, Bangladesh. It is further supported by AI for Earth Grant for a project titled “GWMap: Applying Machine Learning to map groundwater levels in Bangladesh”.

Conflicts of interest/Competing interests

None to be declared.

Availability of data and material (data transparency)

All data are available through the appropriate institutes.

Code availability (software application or custom code)

The code will be made available through github/equivalent repository.

Authors' contributions

Zzaman: Data curation, Formal analysis, Validation, Visualization, Writing - original draft.

Nowreen: Conceptualization, Methodology, Data curation, Formal analysis, Supervision, Writing - review & editing. Khan: Methodology, Formal analysis, Validation.

Islam: Methodology, Formal analysis, Validation.

Ibtehaz: Methodology, Formal analysis, Validation.

Saifur Rahman: Conceptualization, Methodology, Investigation. Anwar Zahid: Data curation, validation.

Farzana: Data curation, validation.

Sharmin: Data curation, validation.

Sohel Rahman: Conceptualization, Methodology, Supervision, Writing - review & editing.

References

Abdollahi, S., Pourghasemi, H. R., Ghanbarian, G. A., & Safaeian, R. (2019). Prioritization of effective factors in the occurrence of land subsidence and its susceptibility mapping using an SVM model and their different kernel functions. *Bulletin of Engineering Geology and the Environment*, *78*(6), 4017–4034. https://doi.org/10.1007/s10064-018-1403-6

Al-Abadi, A. M., Al-Temmeme, A. A., & Al-Ghanimy, M. A. (2016). A GIS-based combining of frequency ratio and index of entropy approaches for mapping groundwater availability zones at Badra–Al Al-Gharbi–Teeb areas, Iraq. *Sustainable Water Resources Management*, *2*(3), 265–283. https://doi.org/10.1007/s40899-016-0056-5

Alam, M. K., Hasan, K. M. S., & Khan, M. R. (1990). Geological map of Bangladesh. United States Geological Survey Bangladesh, Government of Bangladesh. Retrieved from.

Althuwaynee, O. F., Pradhan, B., Park, H.-J., & Lee, J. H. (2014). A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. *CATENA*, *114*, 21–36. https://doi.org/10.1016/j.catena.2013.10.011

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552.

Arabameri, A., Lee, S., Tiefenbacher, J. P., & Ngo, P. T. T. (2020). Novel Ensemble of MCDM-Artificial Intelligence Techniques for Groundwater-Potential Mapping in Arid and Semi-Arid Regions (Iran). *Remote Sensing*, *12*(3), 490. https://doi.org/10.3390/rs12030490

Brammer, H. (2012). *Physical geography of Bangladesh*. The University Press Ltd.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bricker, S. H., Banks, V. J., Galik, G., Tapete, D., & Jones, R. (2017). Accounting for groundwater in future city visions. *Land Use Policy*, *69*, 618–630. https://doi.org/10.1016/j.landusepol.2017.09.018

Burgess, W. G., Shamsudduha, M., Taylor, R. G., Zahid, A., Ahmed, K. M., Mukherjee, A., et al. (2017). Terrestrial water load and groundwater fluctuation in the Bengal Basin. *Scientific Reports*, *7*(1), 3872. https://doi.org/10.1038/s41598-017-04159-w

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Dalin, C., Wada, Y., Kastner, T., & Puma, M. J. (2018). Erratum: Corrigendum: Groundwater depletion embedded in international food trade. *Nature*, *553*(7688), 366–366. https://doi.org/10.1038/nature24664

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).

DPHE. (2008). *Union Wise Water Technology Mapping, Government of the People's Republic of Bangladesh (GoB), Department of Public Health Engineering (DPHE)*. Project Director Sanitation, Hygiene and Water Supply Project (GOB - UNICEF) DPHE, Dhaka.

Earth Resources Observation and Science (EROS) Center. (2021). *U.S. Geological Survey*.

EPC/MMP. (1991). Dhaka region groundwater and subsidence study. Final Report, Vol.2, main report, Dhaka WASA.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Freedman, D. A. (2005). *Statistical models: theory and practice*. Cambridge University Press.

Harvey, C. F., Ashfaque, K. N., Yu, W., Badruzzaman, A. B. M., Ali, M. A., Oates, P. M., et al. (2006). Groundwater dynamics and arsenic contamination in Bangladesh. *Chemical Geology*, *228*(1–3), 112–136. https://doi.org/10.1016/j.chemgeo.2005.11.025

Hasda, R., Rahaman, M. F., Jahan, C. S., Molla, K. I., & Mazumder, Q. H. (2020). Climatic data analysis for groundwater level simulation in drought prone Barind Tract, Bangladesh: Modelling approach using artificial neural network. *Groundwater for Sustainable Development*, *10*, 100361. https://doi.org/10.1016/j.gsd.2020.100361

Hoque, M. A., & Adhikary, S. K. (2020). Prediction of groundwater level using artificial neural network and multivariate time series models. *5th International Conference on Civil Engineering for Sustainable Development (ICCESD 2020),7~9 February 2020, KUET,* *Khulna, Bangladesh*.

Hossain, M, M., & Zahid, A. (2014). Bangladesh Water Development Board: A bank of hydrological data essential for planning and design in water sector. In *Proceedings of the 2nd International Conference on Advances in Civil Engineering 2014. Chittagong University of Engineering and Technology, Bangladesh, December 2014.* Proceedings of the 2nd International Conference on Advances in Civil Engineering 2014. Chittagong University of Engineering and Technology, Bangladesh, December 2014.

Husna, N., Bari, S. H., Hussain, M. M., Ur-rahman, M. T., & Rahman, M. (2016). Ground water level prediction using artificial neural network. *International Journal of Hydrology Science and Technology*, *6*(4), 371–381. https://doi.org/10.1504/IJHST.2016.07934456

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, *14*(1), 1137–1145.

Lee, C. F., Li, J., Xu, Z. W., & Dai, F. C. (2001). Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. *Environmental Geology*, *40*(3), 381–391. https://doi.org/10.1007/s002540000163

Lumley, T., & Lumley, M. T. (2013). Package ‘leaps.’ *Regression Subset Selection. Thomas Lumley Based on Fortran Code by Alan Miller. Available online: http://CRAN. R-project. org/package= leaps (accessed on 18 March 2018)*.

Machiwal, D., Jha, M. K., & Mal, B. C. (2011). Assessment of groundwater potential in a semi-arid region of India using remote sensing, GIS and MCDM techniques. *Water resources management*, *25*(5), 1359–1386.

Miraki, S., Zanganeh, S. H., Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., & Pham, B. T. (2019). Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. *Water Resources Management*, *33*(1), 281–302. https://doi.org/10.1007/s11269-018-2102-6

Nguyen, P. T., Ha, D. H., Nguyen, H. D., Van Phong, T., Trinh, P. T., Al-Ansari, N., et al. (2020). Improvement of Credal Decision Trees Using Ensemble Frameworks for Groundwater Potential Modeling. *Sustainability*, *12*(7), 2622. https://doi.org/10.3390/su12072622

Nowreen, S., Newton, I. H., Zzaman, R. U., Islam, A. K. M. S., Islam, G. M. T., & Alam, M. S. (2021). Development of potential map for groundwater abstraction in the northwest region of Bangladesh using RS-GIS-based weighted overlay analysis and water-table-fluctuation technique. *Environmental Monitoring and Assessment*, *193*(1), 24. https://doi.org/10.1007/s10661-020-08790-5

Oke, S. A., & Fourie, F. (2017). Guidelines to groundwater vulnerability mapping for Sub-Saharan Africa. *Groundwater for Sustainable Development*, *5*, 168–177. https://doi.org/10.1016/j.gsd.2017.06.007

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Qureshi, A. S., Ahmed, Z., & Krupnik, T. J. (2014). *Groundwater management in Bangladesh: an analysis of problems and opportunities*. Cereal Systems Initiative for South Asia Mechanization and Irrigation (CSISA-MI) Project, Research Report No. 2., Dhaka, Bangladesh: CIMMYT.

Razandi, Y., Pourghasemi, H. R., Neisani, N. S., & Rahmati, O. (2015). Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Science Informatics*, *8*(4), 867–883. https://doi.org/10.1007/s12145-015-0220-8

Salam, R., Towfiqul Islam, A. R. M., & Islam, S. (2020). Spatiotemporal distribution and prediction of groundwater level linked to ENSO teleconnection indices in the northwestern region of Bangladesh. *Environment, Development and Sustainability*, *22*(5), 4509–4535. https://doi.org/10.1007/s10668-019-00395-4

Salem, G. S. A., Kazama, S., Shahid, S., & Dey, N. C. (2018). Impacts of climate change on groundwater level and irrigation cost in a groundwater dependent irrigated region. *Agricultural Water Management*, *208*, 33–42. https://doi.org/10.1016/j.agwat.2018.06.011

Sameen, M. I., Pradhan, B., & Lee, S. (2019). Self-Learning Random Forests Model for Mapping Groundwater Yield in Data-Scarce Areas. *Natural Resources Research*, *28*(3), 757–775. https://doi.org/10.1007/s11053-018-9416-1

Sarkar, B. C., Deota, B. S., Raju, P. L. N., & Jugran, D. K. (2001). A Geographic Information System approach to evaluation of groundwater potentiality of Shamri micro-watershed in the Shimla Taluk, Himachal Pradesh. *Journal of the Indian Society of Remote Sensing*, *29*(3), 151–164. https://doi.org/10.1007/BF02989927

Shahid, S. (2008). Spatial and temporal characteristics of droughts in the western part of Bangladesh. *Hydrological Processes*, *22*(13), 2235–2247. https://doi.org/10.1002/hyp.6820

WARPO. (2000). *Draft Development Strategy (DDS), Estimation of groundwater resources, Annex-C, Appendix 6, National Water Management Plan, Water Resources Planning Organization, Ministry of Water Resources*.

Wischmeier, W. H., & Smith, D. D. (1978). *Predicting rainfall erosion losses: a guide to conservation planning*. US Department of Agriculture, Washington, D.C.

Yamamoto, J. K. (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology*, *32*(4), 489–509. https://doi.org/10.1023/A:1007577916868

Zhou, T., Wang, F., & Yang, Z. (2017). Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. *Water*, *9*(10), 781.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.