

Article

# Groundwater Potential Mapping Using Remote Sensing and GIS-Based Machine Learning Techniques

Sunmin Lee <sup>1,2</sup> , Yunjung Hyun <sup>3</sup>, Saro Lee <sup>4,5</sup>  and Moung-Jin Lee <sup>6,\*</sup> 

<sup>1</sup> Department of Geoinformatics, University of Seoul, 163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Korea; smilee@uos.ac.kr

<sup>2</sup> Center for Environmental Assessment Monitoring, Environmental Assessment Group, Korea Environment Institute (KEI), 370 Sicheong-daero, Sejong-si 30147, Korea

<sup>3</sup> Department of Land and Water Environment Research, Korea Environment Institute (KEI), 370 Sicheong-daero, Sejong-si 30147, Korea; yjhyun@kei.re.kr

<sup>4</sup> Geoscience Platform Division, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124, Gwahak-ro Yuseong-gu, Daejeon 34132, Korea; leesaro@kigam.re.kr

<sup>5</sup> Department of Geophysical Exploration, Korea University of Science and Technology, 217 Gajeong-ro Yuseong-gu, Daejeon 34113, Korea

<sup>6</sup> Center for Environmental Data Strategy, Korea Environment Institute (KEI), 370 Sicheong-daero, Sejong-si 30147, Korea

\* Correspondence: leemj@kei.re.kr; Tel.: +82-44-415-7314

Received: 17 February 2020; Accepted: 7 April 2020; Published: 8 April 2020



**Abstract:** Adequate groundwater development for the rural population is essential because groundwater is an important source of drinking water and agricultural water. In this study, ensemble models of decision tree-based machine learning algorithms were used with geographic information system (GIS) to map and test groundwater yield potential in Yangpyeong-gun, South Korea. Groundwater control factors derived from remote sensing data were used for mapping, including nine topographic factors, two hydrological factors, forest type, soil material, land use, and two geological factors. A total of 53 well locations with both specific capacity (SPC) data and transmissivity (T) data were selected and randomly divided into two classes for model training (70%) and testing (30%). First, the frequency ratio (FR) was calculated for SPC and T, and then the boosted classification tree (BCT) method of the machine learning model was applied. In addition, an ensemble model, FR-BCT, was applied to generate and compare groundwater potential maps. Model performance was evaluated using the receiver operating characteristic (ROC) method. To test the model, the area under the ROC curve was calculated; the curve for the predicted dataset of SPC showed values of 80.48% and 87.75% for the BCT and FR-BCT models, respectively. The accuracy rates from T were 72.27% and 81.49% for the BCT and FR-BCT models, respectively. Both the BCT and FR-BCT models measured the contributions of individual groundwater control factors, which showed that soil was the most influential factor. The machine learning techniques used in this study showed effective modeling of groundwater potential in areas where data are relatively scarce. The results of this study may be used for sustainable development of groundwater resources by identifying areas of high groundwater potential.

**Keywords:** groundwater potential; specific capacity; machine learning; boosted tree; ensemble models

## 1. Introduction

Because groundwater has less exposure to pollution than surface water, it is considered a valuable natural resource for agriculture in many communities [1]. Especially during the drought season,

a continuous supply of groundwater is important in agricultural areas. The study area in this investigation, Gyeonggi-do, has recently suffered from damage to agricultural land due to increasing drought. In 2018, widespread damage to crops due to heat waves and drought continued throughout the year, and the average storage rate in 339 reservoirs in Gyeonggi-do was 59% of capacity, which was only 76% of the normal level [2].

Groundwater is a good water resource because it can stably supply the required amount of high-quality water; thus, appropriate water conservation plans are essential for the sustainable use of groundwater [3]. In many areas, the main causes of groundwater depletion are excessive groundwater extraction and unsuitable aquifer recharge [4]. Therefore, accurate estimation and prediction of groundwater recharge should be carried out to support efficient use and systematic management of groundwater resources. From this perspective, groundwater potential mapping using yield data is important. Yield data include extraction volume and the velocity of groundwater at various measurement points. Groundwater yield depends on geological, topographic, and anthropogenic factors specific to the area, and is also related to groundwater potential [5].

In practical terms, groundwater is less accessible than surface water. Groundwater can be presumed by detecting gravity anomalies such as Gravity Recovery and Climate Experiment (GRACE) [6–8]; however, a local groundwater potential map is essential for regional management of groundwater. Thus, studies on the distribution and prediction of groundwater resources have been limited to local scales based on data obtained from point measurements (e.g., meteorological stations, flow measurement points, and groundwater level monitors) [9,10]. In recent years, areal distribution analysis data obtained through remote sensing have been used for global prediction of the water resource distribution in combination with various machine learning techniques, albeit with high uncertainty. To overcome the limitation of groundwater resource surveys based on local information, these data can be converted into global distribution data using satellite imagery. Remote sensing generally produces data in the form of grids or regions, which can be converted into distribution patterns through various processing methods such as machine learning algorithms. By applying the characteristics of remote sensing data to groundwater resources, point-based groundwater hydrological modeling can be extended to the global scale. Therefore, using existing groundwater yield data, it is possible to make regional and local predictions with remote sensing-based methods.

For groundwater potential mapping, a variety of techniques have been applied, including direct drilling for hydrological testing and geophysical models [11,12]. Such methods are suitable for identifying the hydrological characteristics of groundwater, but have high costs in time and money [13,14]. In recent years, studies related to groundwater potential have been conducted using machine learning models with available historical data on groundwater wells with geographic information systems (GIS) [15,16]. GIS technologies have been used for quantitative analysis of spatial distributions in environmental, geological, and hydrological studies [17–19]. One limitation of data-based analysis of groundwater is insufficient availability of data for analysis [20]; groundwater yield varies with hydrological conditions and recharge sources, which have been measured in a limited number of groundwater wells [21]. Therefore, using various models to predict groundwater yield accurately and identifying the optimal model for water resource evaluation in a given region are essential to effective water resource management.

For this reason, studies related to groundwater potential mapping with various data models have become increasingly common [22–24]. Numerous factors that affect groundwater potential have been proposed based on various data modeling methodologies, including statistical models, probabilistic models, machine learning models, and data mining models; yield and spring or well location data are also widely used as groundwater potential indicators. Due to the characteristics of remote sensing and groundwater, groundwater could be indirectly monitored by using remote sensing; much research has been conducted through thematic maps related to groundwater based on remote sensing data and groundwater potential was estimated by reducing the uncertainties [25–27].

The frequency ratio (FR) model is a representative statistical model applied to groundwater potential mapping [26,28,29]. The relationship between groundwater conditioning factors and groundwater potential could be analyzed using basic statistical and probabilistic models, including FR, weight of evidence [30], evidential belief function [31], and logistic regression [32] models. Furthermore, the recent exponential increase in available data has led to identification of data types and data processing techniques that can support decision-making. Several studies in this area have applied machine learning methods such as machine learning models, while artificial neural networks [33] and support vector machines [34] have been widely applied to groundwater potential mapping. Some studies have also used analytical hierarchy methods, which are expertise-based methods requiring a deep understanding of the study area [35,36]. Recently, hybrid and ensemble models that combine or develop existing methodologies have been applied for groundwater potential mapping [37–39]. This paper also uses a hybrid methodology in this respect.

When performing groundwater potential mapping through modeling, the results show poor generalizability without proper training samples. In such cases, the accuracy for training data is high but the testing results show significantly lower accuracy. To overcome the lack of data, robust models built upon basic models have recently been developed and compared [40]. Typically, an ensemble model using learner sequences is developed; voting, bagging, and adaptive boosting are representative ensemble methods that can be applied to various base learners [41]. In this way, unlabeled cases are identified via self-learning by combining information from labeled cases so that the labeled training set is magnified in each iteration until the entire dataset is labeled. This method, which was applied in the present study, could be effective for data-scarce areas because it allows modeling using less data than other approaches.

Previous studies conducted on groundwater recharge and yield have used enough field survey data targeted at adjacent areas. However, these studies are subordinate to field surveys and are not intended to reduce spatial uncertainty on groundwater. Therefore, the purpose of this study was to map and test groundwater yield potential in Yangpyeong-gun, South Korea, using spatial data analysis in a GIS environment. This study processed and analyzed officially published groundwater yield data using remote sensing and GIS to reduce the uncertainty of the data itself. In addition, one of the latest machine learning models, boosted tree method, was applied to predict large areas of low uncertainty using pumping test data from 53 wells; groundwater yield potential is the major issue of this study. The results of this study could provide a scientific basis for efficient use and systematic management of groundwater resources.

## 2. Study Area

South Korea consists of eight administrative districts, labeled ‘-do’, which are made up of local administrative districts, labeled with ‘-si’, ‘-gun’, and ‘-gu’. The study area, Yangpyeong-gun, is located about 50 km from Seoul, in the northeastern part of Gyeonggi-do (Figure 1). Yangpyeong-gun is surrounded by Hongcheon-gun in Gangwon-do to the northeast, Hoengseong-gun in Gangwon-do to the east, Wonju-si in Gangwon-do to the southeast, and Gapyeong-gun to the north. Yangpyeong-gun contains rugged mountainous areas such as Yongmunsan (1157 m), Bongmyun (856 m), and Baekunbong (940 m), and the Namhan River flows from the south to the northwest of the district. About 90% of the total area of Yangpyeong-gun is a green zone covering the protected headwater area of the Han River; this area has a well-preserved and clean natural environment due to legal and institutional regulations [42].

Yangpyeong-gun covers approximately 878 km<sup>2</sup>, and the amount of groundwater used in this area is 41,503,946 m<sup>3</sup>/year. The groundwater use per unit area is 47,258 m<sup>3</sup>/km<sup>2</sup> annually and 129 m<sup>3</sup>/km<sup>2</sup> daily [43]. Groundwater in Gyeonggi-do is used primarily for agricultural purposes in numerous agricultural areas, including Anseong-si, Yangpyeong-gun, Icheon-si, and Yeosu-si. Among all districts in South Korea, Yangpyeong-gun (10,725) has the second highest number of groundwater facilities for agricultural use after Anseong-si [43].

In Yangpyeong-gun, a preliminary survey of available groundwater resources was conducted from December 2017 to June 2018 for drought response and to prevent unplanned development. Among 35 districts prone to drought, 25 were selected based on the feasibility of surveying the target district and the response rate of residents. A resistivity survey (vertical and dipole survey) was conducted to select locations for large-scale groundwater storage.

In Yangpyeong-gun, Gyeonggi-do, Kyonggi massif metamorphic rocks of Precambrian age and an intrusive body of Mesozoic Triassic gabbro and syenite are found. Precambrian Kyonggi massif metamorphic rocks consist of the Paleozoic sequence of Yongmunsan and unconformity of Jang-Rak. The main constituent rocks are banded gneiss, migmatitic gneiss, augen gneiss, mica schist, and quartzite. These rocks underwent metamorphism in the Paleozoic and Mesozoic Triassic, when the landmasses of North China and South China collided.

Groundwater development requires continuous management for sustainable supply of water rather than short-term measures at the time of drought. Specifically, preliminary investigation is needed in drought-prone areas and areas of high importance for agricultural water usage in Gyeonggi-do. To mount an effective response to agricultural drought, a groundwater management plan that ensures sustainable use of agricultural groundwater prior to drought is needed [44]. In this study, continuous groundwater potential data in the study area were used as primary data for a groundwater abundance survey, and could further be used to establish a groundwater development plan.

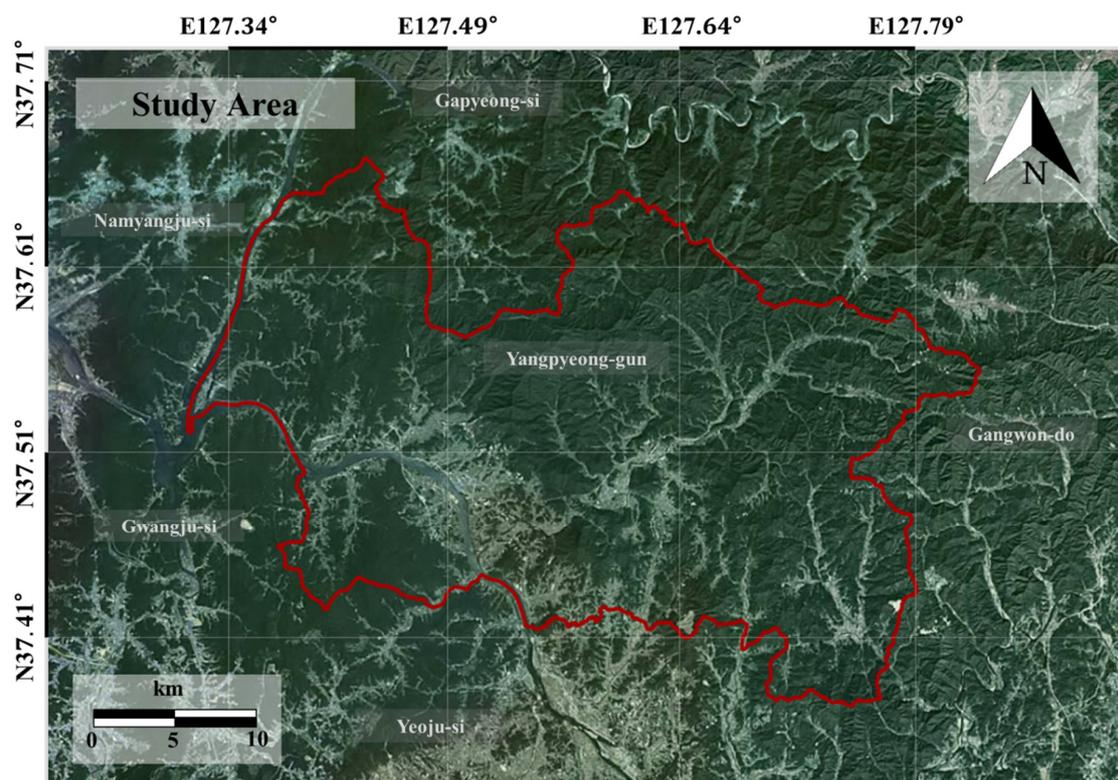


Figure 1. Study area.

### 3. Data

#### 3.1. Groundwater Potential Analysis Based on Remote Sensing Data

Various thematic maps constructed using remote sensing source data were applied to machine learning techniques in this study. Recently, high-resolution aerial photographs were used to produce thematic maps of spatial data. Topographic maps were produced through numerical mapping using aerial photographs taken in 2006, with corrections and supplemental data collected through field surveys. Forest and soil maps were also constructed using spatial data generated through field surveys

along with aerial photography. For land use maps, aerial photographs taken in 2012 were classified using image classification techniques, and their quality was verified using additional high-resolution satellite images from KOMPSAT-2 and KOMPSAT-3 as well as digital topographic maps. Meanwhile, geological maps were produced from field surveys and historical records using base maps generated from aerial photographs. Groundwater yield is a measure of groundwater pumping capacity, which could be stored in aquifers. In this study, groundwater yield potential modeling using machine learning was performed with spatial data generated via remote sensing and GIS such as soil, land cover, and geological maps, as described above.

### 3.2. Groundwater Well Data from *in Situ* Sampling

Groundwater pumped from wells in the study area is used mainly for agricultural purposes and domestic drinking water. Groundwater well data were collected for specific capacity (SPC) (53 wells) and transmissivity (T) (53 wells) from the basic survey report of Yangpyeong-gun [45]. The main use of the groundwater in this area is agricultural, so groundwater surveys are conducted between spring and summer, and our data was obtained between June and August. In the training and testing subsets, yield values above 3.8 and 3.42 (30 m<sup>3</sup>/h) above the median value were considered for yields based on the dependent variables of SPC and T, respectively, which are two different indexes measured in different ways. Groundwater pumping test data used in this study were generated and published from the national groundwater observation and survey data by local governments conducted by Korea Water Resources Corporation (K-water).

SPC data include geographic location coordinates of individual wells and groundwater yield derived from pumping tests. SPC often indicates well performance, because it refers to the amount of water that a well can produce per unit of drawdown. SPC is calculated by dividing the pumping discharge by the drawdown, in units of liters per minute (LPM) per meter, as follows:

$$SPC = \frac{Q}{S} \quad (1)$$

where Q is discharge (unit: LPM) and S is drawdown (unit: m). A low SPC value indicates that more energy is required for pumping. During a drawdown test to determine SPC, pumping should be maintained at a constant speed for a certain period of time, at least 24 h, with little change in drawdown. SPC data acquired during the pumping test can be used to estimate T and identify potential aquifer issues.

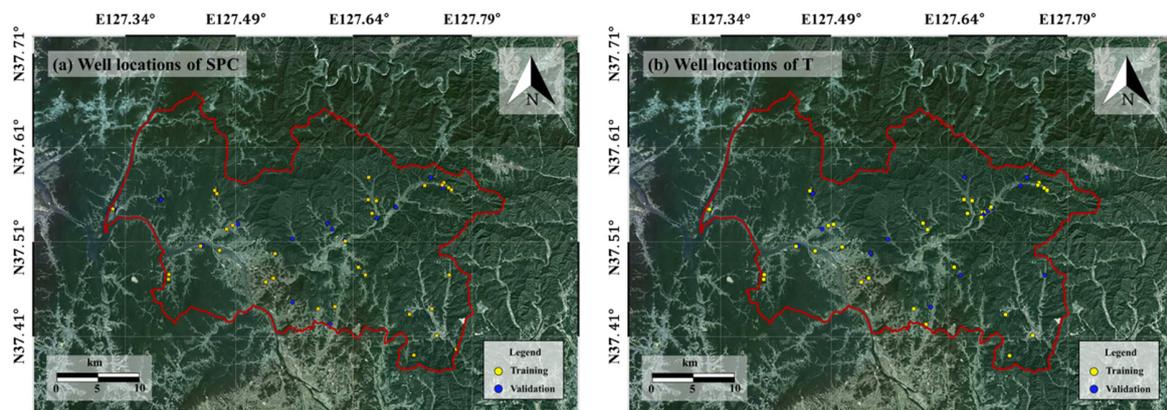
T represents the flow rate under a unit hydraulic gradient through a unit width of aquifer of a certain thickness [46]. Hydraulic conductivity (K) is a measure of the water transmission capacity of an aquifer. T of an aquifer is equal to the hydraulic conductivity multiplied by the thickness of the aquifer.

$$K'(x, y) = \frac{1}{b} \int_0^b K(x, y, z) dz \quad (2)$$

$$T = Kb, \quad (3)$$

where T is transmissivity, K is hydraulic conductivity, and b is aquifer thickness. Less drawdown and a thicker aquifer lead to higher T values. It is possible to estimate the amount of water flowing through the unit thickness of the aquifer by combining Equation (3) with Darcy's law.

SPC and T data were separately applied to the FR, boosted tree (BT), and ensemble models in this study; both SPC and T are used in this study in order to consider various aspects of groundwater. The locations of groundwater wells in the study area are shown in Figure 2. Yield data were randomly divided into a training data subset (70%) and a testing data subset (30%), as is the usual division in machine learning methodologies [16,47]. In the training data subset, 37 wells each were represented in SPC and T data, respectively; 16 wells were used to test the models.



**Figure 2.** Locations of groundwater wells sampled for: (a) SPC and (b) T data.

### 3.3. Groundwater Conditioning Factors

Various groundwater conditioning factors were used for groundwater potential modeling in this study (Table 1). Topographical, geological, hydrological, and land cover factors are commonly applied to predict groundwater yield potential. Conditioning factors should be considered depending on regional characteristics. For this reason, the correlation between the factors and groundwater potential were analyzed preferentially through the frequency ratio model and the factors were selected; groundwater potential was estimated using 16 factors in this study. The 16 conditioning factors were constructed into a groundwater inventory, including nine topographic factors (convergence index, convexity, mass balance index (MBI), slope angle, slope height, topographic texture, topographic position index (TPI), topographic ruggedness index (TRI), and valley depth), two hydrological factors (flow path length, and slope length and steepness (LS)), forest type, soil material, land use, and two geological factors (lithology and distance from fault) (Figures 3 and 4). The conditioning factors were calculated and prepared using ArcGIS 10.3 software (ESRI, Redlands, CA, USA). Each dataset was converted into a grid format with 30-m spatial resolution for use in the groundwater inventory of the study area.

Topographic factors were calculated from a 1:5000 scale topographic map provided by the Korean National Geographic Information Institute. Spatial data, such as location and topography, were structured using ground control point measurements taken from digital aerial photographs and ground surveys. Aerial photographs were analyzed through numerical mapping, and further calibration was carried out through field surveys to create the topographic map. A digital elevation model (DEM) was first generated from the topographic map and then used to derive topographic factors, including convergence index, convexity, MBI, slope angle, slope height, topographic texture, TPI, TRI, and valley depth. Slope factor impacts groundwater recharge, with gentle slope areas having relatively high percolation and low surface runoff rates and steep areas having high surface runoff [48]. Soil moisture content is also related to slope, which affects precipitation direction [49]. Slope angle is strongly related to groundwater potential; therefore, groundwater-related topographic factors derived from DEM data with SAGA-GIS software [50] were used for modeling. Acceleration and deceleration, as well as flow convergence and divergence of flow, are mainly affected by the curvature of the area [51]. The hydrological factors flow path and LS factor were considered conditioning factors for hydrological features.

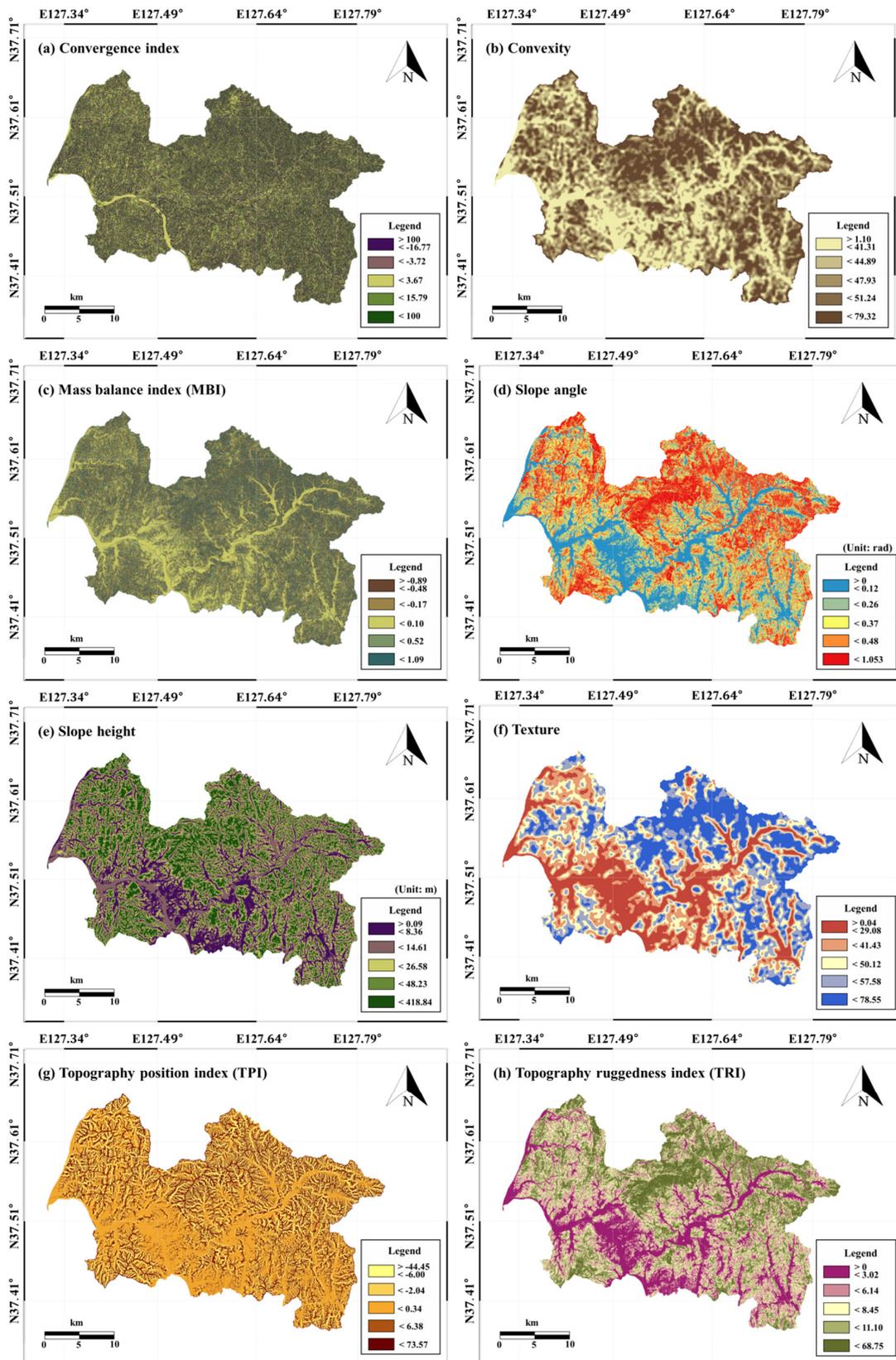
A forest map was also used, which was generated from field investigations and interpretation of aerial photographs. To construct the forest map, the near-infrared band was used for image analysis, in addition to the red-green-blue image. Moreover, soil material characteristics can impact the rate of surface water penetration into aquifers, which drives groundwater potential [52]. The soil material factor was extracted from a soil map published by the National Institute of Agricultural Sciences at 1:25,000 scale. Similarly, land cover has an impact on soil conditions such that storage and movement

of groundwater change when land cover changes; the land use factor was extracted from a digital land cover map provided by the Korea Ministry of Environment at 1:25,000 scale. Land use maps were classified into 22 medium-level categories through application of automatic image classification to aerial photographs, and the accuracy was enhanced using additional high-resolution satellite images from KOMPSAT-2 and 3. The land cover map was reclassified into seven land cover categories: urban, farmland, forest, grassland, wetland, bare land, and water.

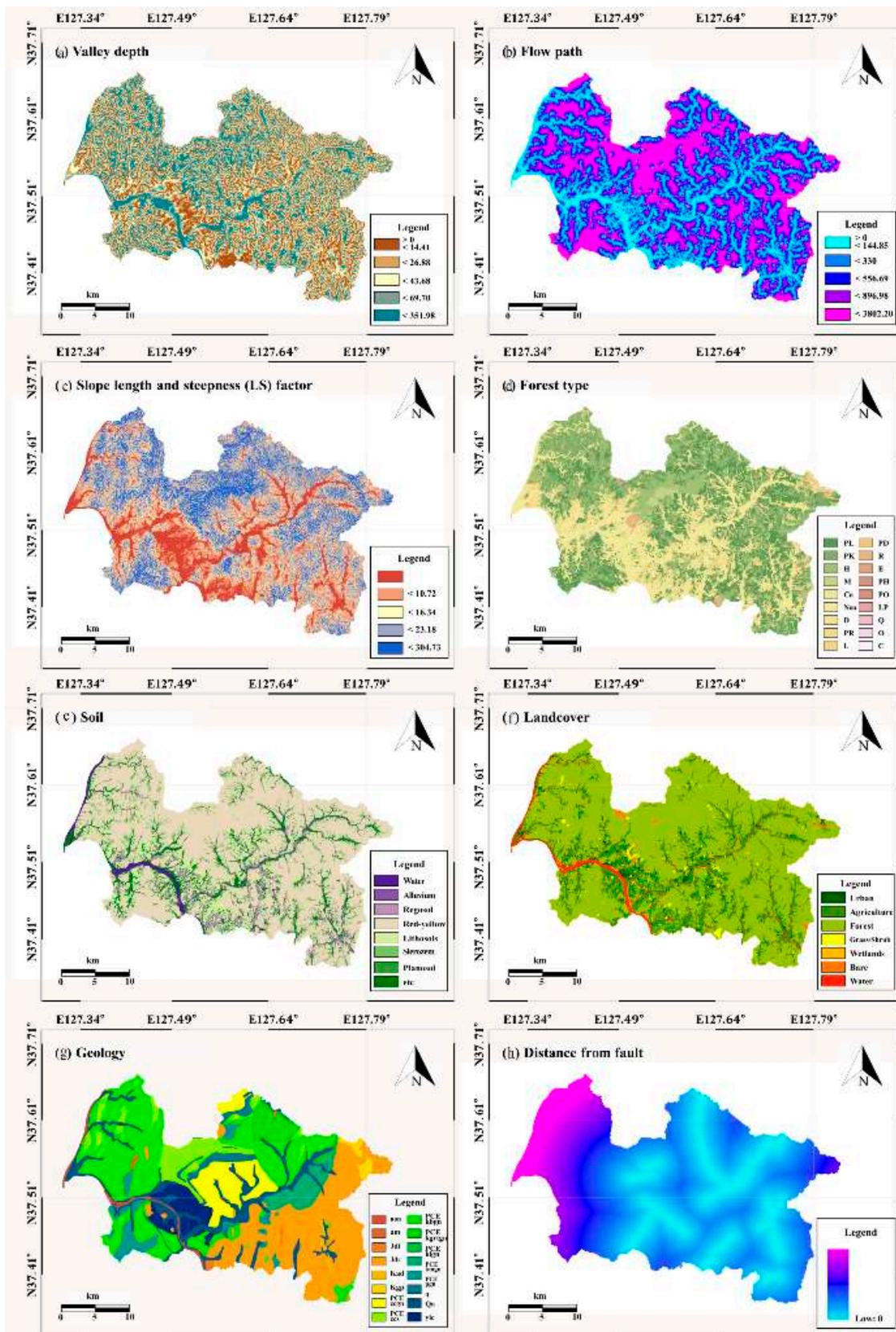
Geological factors, including lithology and distance from a fault, were also considered in relation to groundwater characteristics. The lithology factor was extracted from a digital geological map produced by the Korea Institute of Geoscience and Mineral Resources at 1:50,000 scale. The study area was composed of 22 lithological units differing in lithology type and geological age. Distance from a fault was also calculated based on the geological map.

**Table 1.** Data layers describing groundwater potential.

Category	Factor	Scale	Data Type	Source Data (Year)
Pumping Test data	Specific capacity (SPC) Transmissivity (T) Convergence index Convexity Mass balance index (MBI)	-	Point	Field Survey (2008)
Topography	Slope angle Slope height Topographic texture Topographic position index (TPI) Topographic ruggedness index (TRI) Valley depth			Aerial Photography (2006–2016)
Hydrology	Flow path Slope length and steepness (LS) factor	1:50,000	Polygon	Field Survey (2008)
Forest	Forest type	1:25,000	Polygon	Aerial Photography, Field Survey (2004–2006)
Soil	Soil	1:25,000	Polygon	Aerial Photography, Field Survey (1998–2006)
Landcover	Landcover	1:5000	Polygon	Kompsat 2, 3 and Aerial Photography (2012)
Geology	Geology Distance from fault	1:250,000	Polygon	Aerial Photography, Field Survey (2004)



**Figure 3.** Groundwater conditioning factors I: (a) Convergence Index, (b) Convexity, (c) Mass balance index (MBI), (d) Slope angle, (e) Slope height, (f) Texture, (g) Topography position index (TPI) and (h) Topography ruggedness index (TRI).



**Figure 4.** Groundwater conditioning factors II: (a) Valley depth, (b) Flow path, (c) Slope length and steepness (LS) factor, (d) Forest type, (e) Soil, (f) Land cover, (g) Geology and (h) Distance from fault.

### 4. Methodology

To be more specific, the purpose of this study was to map and test groundwater yield potential in Yangpyeong-gun, South Korea, using spatial data analysis in a GIS environment. This was performed by four main steps: First, groundwater yield data of specific capacity (SPC) and transmissivity (T) collected from 53 well locations were used. For the training data, 70% of each groundwater yield dataset was selected randomly, and FR and boosted tree (BT) models with classification were applied to the groundwater inventory using Statistica software (Dell Software, Aliso Viejo, CA, USA). Second, the inventory was constructed from nine topographic factors, two hydrological factors, forest type, soil material, land use, and two geological factors. All factors used in this study were generated and processed from remote sensing-based data, such as aerial photographs or imagery from KOMPSAT-2 and -3. Third, this study involved probabilistic analysis of FR, and two machine learning models: the boosted classification tree (BCT) and FR-BCT ensemble models, which were applied to groundwater yield data. Comparative analysis was conducted to compare the models used in this study. Finally, to quantitatively evaluate the performance of the models, the receiver operating characteristics (ROC) and area under the curve (AUC) were used. The study was conducted, as shown in Figure 5.

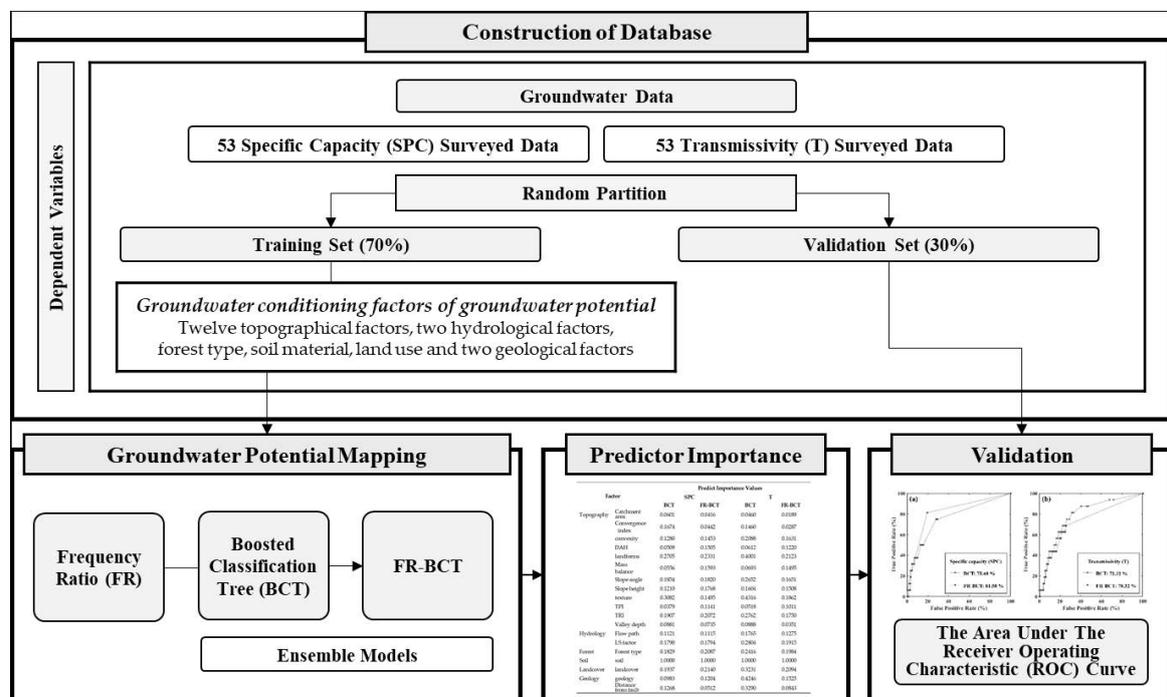


Figure 5. Data flow of this study.

#### 4.1. Frequency Ratio (FR) Model

FR is an effective stochastic method for evaluating the effects of various factors on the occurrence of a particular event [53]. Thus, the FR value represents the ratio of occurrence of a particular event to the area ratio for each class [54]. A larger FR value represents a stronger relationship between the probability of occurrence and the specific variable [55,56]. This method allows for the clear and simple analysis of the relationship of each factor to the event [57].

To carry out spatial FR analysis, factors related to groundwater potential were classified into ten classes. Among numerous available classification techniques, factors in this study were classified using the quantile technique, which divides classes into equal areas. FR values were calculated using training data for each factor. Each class of each modulator was weighted. Higher FR values represent a stronger relationship between the class of each factor and groundwater potential, whereas for lower FR values, the effect of the class of each factor on groundwater potential is small. If FR is greater than 1, the effect is significant; if FR is less than 1, the effect is not significant [56]. To construct a groundwater

potential map using FR to represent the relative magnitude of the groundwater potential, the FR values calculated for each factor were determined as follows:

$$FR = \frac{P_{trn}}{P_{total}} \quad (4)$$

where  $P_{trn}$  is the ratio of the number of SPC data points above a certain level and  $P_{total}$  indicates the ratio of the number of pixels in a certain class to the total number of pixels in the study area. A greater FR value for potential indicates higher groundwater potential; a lower value indicates a lower groundwater potential. In this study, FR values for each conditioning factor were used to weight the ensemble FR-BCT model.

#### 4.2. Boosted Classification Tree

In recent years, decision tree models have been used in various fields as a machine learning method [58], including for groundwater potential mapping [52]. Decision tree models perform attribute tests on non-terminal nodes to represent the results on the terminal node, using a tree-like hierarchy that constructs a classification tree of a simple structure [59]. One of the benefits of this method is that the classification process can be graphically represented. However, the results cannot be formed into multiple outputs and the performance of the model depends on the type of data. Many algorithms have been developed from decision trees: classification and regression tree [60], chi-square automatic interaction detector decision tree [61], Iterative Dichotomiser 3 [62], and J48 (C4.5 decision tree) [63]. In addition, ensemble models using sequences of classifiers have been widely developed. Representative ensemble methods such as voting, bagging (sub-sampling), and boosting have also been applied to the decision tree method, including BT algorithms. Therefore, in this study, representative decision tree algorithms of BT models were used to compare the performance of each model's groundwater potential modeling and prediction accuracy.

The BT model is a tree-based machine learning model using the stochastic gradient boosting method. In the last few years, this algorithm has become one of the most powerful machine learning techniques used for prediction. In the BT algorithm, continuous or categorical input factors can be used for classification and regression problems [64].

The BT algorithm is implemented by applying a boosting method to the regression tree. The basic method involves calculating a simple tree sequence in which each successive tree is built against the prediction residual of the preceding tree. This method creates two trees of data for two samples at each split node. Even if the relationship between predictive and dependent variables is nonlinear, the weighting of such trees can support high accuracy of the predicted value. Thus, the gradient boosting method for weighted expansion of simple trees is one of the most common and powerful machine learning algorithms.

All machine learning algorithms are prone to overfitting, which involves a good fit for learning data but a lack of improvement in the predictability of each model. In other words, this is a common problem that applies to most algorithms used for predictive machine learning. A common solution to this problem is to evaluate the quality of the model fit by predicting observations from test samples of "used" data before evaluating each model [65,66]. The accuracy of each solution can be measured in this way to determine when the overflow occurred.

To overcome this difficulty, which is a major problem facing most machine learning algorithms used in predictive models, a specific approach was selected for the BT models. A continuous simple tree is generated using only subsamples selected randomly from the entire dataset. That is, each successive tree is created for the predicted residuals of an independently extracted random sample. Randomness can be added to any degree to protect against overfitting and can provide good predictability. Continuous boosting calculations for independently sampled input samples are known as probabilistic gradient boosting techniques.

### 4.3. Ensemble Modelling

Using the two methodologies described above, ensemble methods of FR and BCT were applied in this study. The probabilistic method FR was used to assess the impact of all types of regulatory factors and assign appropriate weights to each class according to their impact on groundwater yield. Using the FR method, individual weights were derived for each factor. Each conditioning coefficient was then reclassified using the derived weight values, and the reclassified dataset was analyzed using the BCT tree-based machine learning models. Finally, a groundwater potential map was constructed using the BCT and FR-BCT ensemble techniques for comparative analysis.

### 4.4. Assessment on Model Performance

The performance of groundwater potential classification was assessed using two statistical indicators: sensitivity and specificity. Sensitivity is the percentage of correctly classified pixels in areas with high groundwater potential; specificity is the percentage of pixels classified as having a low groundwater potential. Sensitivity and specificity are calculated as follows [67]:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (6)$$

The numbers of correctly classified pixels are denoted as true positives (TP) and true negatives (TN). Conversely, the numbers of misclassified pixels are expressed as false positives (FP) and false negatives (FN).

In this study, ROC curves were used to evaluate the overall performance of the groundwater potential model. The ROC curve has been applied in various fields as a standard method for evaluating the general performance of a model [68]. This curve is plotted using sensitivity as the x-axis and 100 – specificity as the y-axis. The general performance of the model can be quantitatively assessed based on the AUC value, representing the area under the ROC curve. AUC values range from 0.5 to 1. A value of 0.5 represents a model with very low accuracy. In contrast, 1 represents a perfect model with the highest possible accuracy, and an AUC close to 1 indicates good performance. Generally, when the AUC value is greater than 0.8, the model shows adequate performance [69].

## 5. Results

### 5.1. Results from the Frequency Ratio Model

Table A1 presents the correlations of FR values between groundwater data (SPC or T) and groundwater conditioning factors derived from the FR model. The FR is a representative value of the statistical proportional position of well locations with SPC values above a specific level. Correlation between groundwater well data and each factor could be shown from the distribution of values biased according to each class. Areas with high FR values are of great importance for groundwater management because they have high groundwater potential. The characteristics of land cover in the area of this study are high in forest area and agricultural area, and relatively low in urban area. Although there are many groundwater wells in urban areas, the urban area is mixed with rural areas, so it requires a different approach from metropolis.

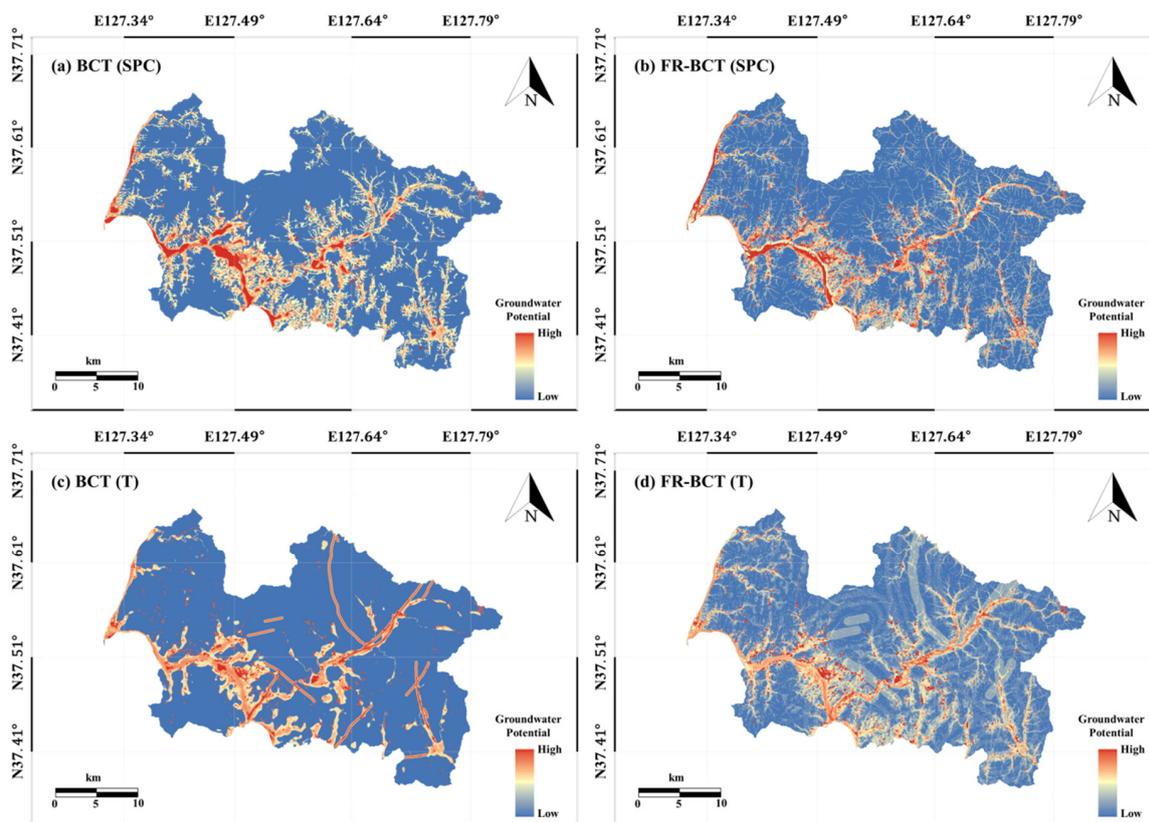
The topographic factor convexity showed a strong correlation with groundwater potential in the 1.1–43.19 class for FR values of over 1.89 and 2.63 for SPC and T, respectively. Similarly, MBI showed a high correlation with SPC (2.16) and T (1.84) in the -0.33 to 0.1 class. The highest FR values of 4.32 for SPC and 4.21 for T were observed when the slope angle was greater than 0 m and less than 0.05 m, indicating that this factor is strongly correlated with groundwater potential. FR values tended to decrease with increasing slope angle and slope height. For topographic texture, the 0.04–29.08 class

exhibited the highest FR values with SPC (2.97) and T (3.95). Low flow path values also led to FR values over 1, indicating that this factor was correlated with groundwater potential.

Among land cover types, urban area showed the strongest relationship with groundwater potential (SPC: 6.66; T: 7.92), followed by wetlands. These results could also be interpreted as showing that the use frequency of wells in urban areas is high. Meanwhile, distance from a fault had FR values of 2.16 for SPC and 3.16 for T in the 0–530.75 class. Among geological factors, alluvium showed a strong correlation with the groundwater data (SPC: 2.93; T: 3.80), followed by granite porphyry (SPC: 1.45; T: 1.01).

## 5.2. Construction of Groundwater Potential Maps

The groundwater potential map was modeled using training datasets of SPC and T. The performance of a groundwater potential model depends on the selection of factors. The groundwater potential map was constructed by training the groundwater potential model. First, a groundwater potential value was generated for each pixel in Yangpyeong-gun. Each pixel was indexed by its predicted groundwater potential value. The results of groundwater potential were reclassified using the 1.0 standard deviation method, which is based on the distribution of individual values in the results for each model. In the groundwater potential map, areas with high (low) groundwater potential are shaded red (blue) (Figure 6). All models showed similar distributions of groundwater potential, and the north, southwest, and southeast areas surrounding the central valley region of the study area all showed low potential.



**Figure 6.** Groundwater potential maps based on (a) boosted classification tree (BCT) and (b) frequency ratio (FR)-BCT models with specific capacity (SPC) data, and (c) BCT and (d) FR-BCT models with transmissivity (T) data.

Furthermore, the predictor importance values of each factor were calculated from the BCT modeling results by summing the decreases in node-impurity values (Table 2). All predictor importance values

were scaled to a maximum of 1.0, as the value assigned to the largest sum among all factors, indicating the most strongly related factor, relatively. For both SPC and T, soil showed the highest predictor importance values in all models, with a value of 1.0. Topographic texture was the second most important factor in the BCT models, with values of 0.3101 and 0.4206, for SPC and T data, respectively. Meanwhile, FR-BCT models showed that forest type and land cover were the second strongest predictors, with importance values of 0.1704 and 0.2295 for SPC and T data, respectively. The importance of TPI, MBI, and valley depth were low in all FR models; convergence index, valley depth, and distance from a fault fell into the third lowest positions based on the FR-BCT models.

**Table 2.** Predictor importance values of each factor for the BCT and FR-BCT models.

Factor		Predictor Importance Values			
		SPC		T	
		BCT	FR-BCT	BCT	FR-BCT
Topography	Convergence index	0.1689	0.0400	0.1518	0.0494
	convexity	0.1285	0.1223	0.1913	0.1698
	Mass balance index (MBI)	0.0566	0.1345	0.0703	0.1680
	Slope angle	0.1909	0.1443	0.2734	0.1850
	Slope height	0.1245	0.1393	0.1711	0.1750
	Topographic texture	0.3101	0.1196	0.4206	0.1975
	Topographic position index (TPI)	0.0387	0.0990	0.0565	0.1246
	Topographic ruggedness index (TRI)	0.1967	0.1658	0.2917	0.2003
	Valley depth	0.0887	0.0696	0.0929	0.0513
Hydrology	Flow path	0.1123	0.0917	0.1819	0.1407
	Slope length and steepness (LS) factor	0.1835	0.1466	0.2747	0.1935
Forest	Forest type	0.1821	0.1704	0.2084	0.1694
Soil	Soil	1.0000	1.0000	1.0000	1.0000
Landcover	Landcover	0.1946	0.1572	0.3285	0.2295
Geology	geology	0.1002	0.0996	0.1619	0.1386
	Distance from fault	0.1271	0.0497	0.3105	0.1061

### 5.3. Model Performance Evaluation

In this study, the groundwater potential model was evaluated based on statistical indices; AUC was used to quantitatively assess the mapping accuracy. As aforementioned, testing was performed based on the 30% of the groundwater well data collected by field investigation; and since groundwater has less seasonal change than surface water, this study did not consider seasonal change for groundwater. Figure 7 presents the model accuracy rate for the SPC (BCT model: 80.48%; FR-BCT model: 87.75%) and T (BCT model: 72.27%; FR-BCT model: 81.49%) well data. In general, all groundwater potential mapping results and modeling of groundwater potential showed good performance; however, the ensemble models showed improved accuracy by approximately 6%. Figure 7 also shows the performance of the groundwater potential models using the ROC curve method. All groundwater potential models performed well in terms of groundwater potential evaluation results ( $AUC > 0.7$ ). The testing results of the BCT ensemble model show that 20% of the groundwater potential area includes approximately 80% of the valid groundwater wells for SPC, whereas the testing results of the ensemble model for T show that 30% of the groundwater area includes over 80% of the valid groundwater wells. Compared to groundwater potential mapping with the single machine learning model, BCT, all groundwater potential models using the ensemble method with both FR and BCT showed better performance, with 7.27% and 9.22% higher accuracy, respectively, than the BCT model alone. The difference in AUC results showed that the ensemble model provided better results than the individual modeling process.

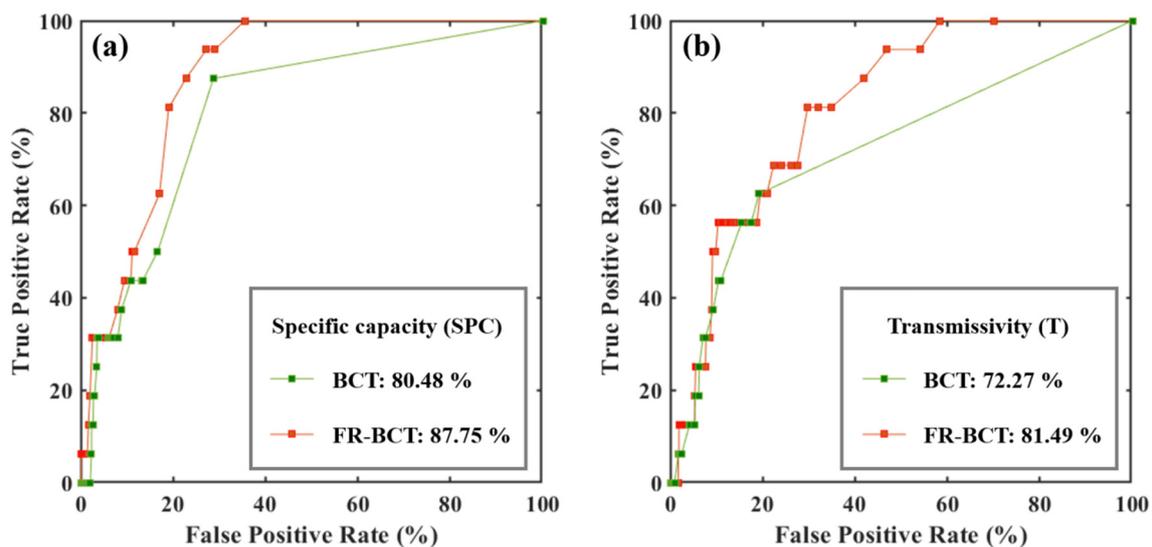


Figure 7. Testing results of the BCT and FR-BCT models for (a) SPC and (b) T groundwater data.

## 6. Discussion

In this paper, the relationship between conditioning factors and groundwater was first analyzed through the stochastic method of FR. By applying the ensemble technique to the BCT model based on the stochastic weighting, it showed effectiveness in the study of groundwater with high uncertainty. In terms of data, this study was based on data created by governments and public institutions and released to the public; at the same time, it is bound by limitations in data collection. Since the importance of data used for training in data-based learning is very high, model accuracy will be improved if more well data is used in future studies.

Few case studies have applied ensemble models from machine learning algorithms in South Korea. The results of this study confirm that the performance of a groundwater potential model can be improved using an existing probability model and machine learning ensemble. Model performance was evaluated based on the ROC, and the prediction rate of the BCT model showed an improvement of 6.1% with FR-BCT for SPC and 6.0% for T compared to the single machine learning model, BCT, indicating that the ensemble method greatly improved model performance. This improvement occurred because the ensemble model could reduce bias using the BT model and improve its predictive ability by avoiding the overfitting problem of basic classification [70]. This finding is consistent with other studies that concluded that the predictive performance of models was improved with a machine learning ensemble model [71].

Remote sensing is a powerful data source that is widely used for monitoring environmental issues; however, since groundwater does not exist on the surface, groundwater can only be indirectly estimated by using remote sensing. Heretofore, many studies have attempted to reduce the uncertainty of groundwater spatially. As a result of applying the proposed FR-BCT model with existing probability models and the machine learning method of the BCT model, the accuracy was relatively improved or similar to previous studies [3,25,34,68]. In addition, by showing accuracy improvements in single and composite models, it has shown potential for reducing the uncertainty of groundwater potential mapping.

## 7. Conclusions

The modern global water shortage requires effective water management and planning. Indiscreet use of water resources and inadequate water management can disrupt the continuous and reliable supply of water. The first step in properly planning water resource usage is to accurately predict and respond to the current status of critical resources. Groundwater represents an excellent water source, especially in water-scarce regions. However, the uncertainty of groundwater availability is

high; therefore, estimation of groundwater potential is essential. Mapping of groundwater potential is an essential challenge facing effective groundwater resource management and conservation planning.

Various methods of groundwater potential mapping have been proposed. Improvement of the groundwater potential model is one method for estimating the uncertainty of a groundwater model. Although new machine learning technologies are continually improving in predictive performance, not all methods can be effectively applied in areas where data are scarce, because it may not be possible to generalize from a small labeled dataset. Therefore, FR analysis and the BCT model were applied along with the proposed FR-BCT model, which is an ensemble model of these two machine learning models. For this purpose, 16 groundwater control factors based on remote-sensing data were applied to the models: nine topographic factors, two hydrological factors, forest type, soil material, land use, and two geological factors. The model was trained and tested using groundwater well data; 53 wells were separated into training (70%) and testing (30%) datasets. The proposed FR-BCT model was compared with existing probability models and the machine learning method of the BCT model.

These results are useful for supporting comprehensive management of groundwater exploration and groundwater recharge. The method used in this study can be applied to other areas reliant on groundwater use. Managers and policymakers can effectively analyze groundwater potential modeling results to maximize the benefits of management. However, further testing is required in other research areas to determine how reliably the proposed ensemble model reflects groundwater potential.

**Author Contributions:** Conceptualization, S.L. and M.-J.L.; Data curation, S.L. and S.L.; Formal analysis, Y.H.; Investigation, Y.H. and S.L.; Methodology, S.L. and Y.H.; Project administration, M.-J.L.; Resources, S.L.; Software, S.L. and S.L.; Supervision, M.-J.L.; Validation, S.L.; Visualization, S.L.; Writing—original draft, S.L.; Writing—review and editing, M.-J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was conducted at the Korea Environment Institute (KEI) with support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041203). This research was conducted by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM) funded by the Ministry of Science and ICT. This research was also conducted with support from ‘Future Forecasting Government Management Support: A Study on the Modeling of Climate and Environment for the Transition to Data Science’ funded by the National Research Council for Economics, Humanities and Social Sciences.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Table A1.** Results of the frequency ratio model.

Factor	Class	No. of SPC	% of SPC	No. of T	% of T	No. of Pixels in Domain	% of Pixels in Domain	Frequency Ratio of SPC	Frequency Ratio of T
Convergence index	−100−−29.9	7	18.92	6	16.22	97,261	10.00	1.89	1.62
	−29.9−−16.77	2	5.41	2	5.41	97,262	10.00	0.54	0.54
	−16.77−−8.98	5	13.51	4	10.81	97,261	10.00	1.35	1.08
	−8.98−−3.72	5	13.51	5	13.51	97,262	10.00	1.35	1.35
	−3.72−0	5	13.51	3	8.11	89,697	9.22	1.47	0.88
	0−3.67	4	10.81	6	16.22	104,826	10.78	1.00	1.50
	3.67−8.59	2	5.41	2	5.41	97,261	10.00	0.54	0.54
	8.59−15.79	2	5.41	4	10.81	97,262	10.00	0.54	1.08
	15.79−28.28	2	5.41	3	8.11	97,261	10.00	0.54	0.81
	28.28−100	3	8.11	2	5.41	97,262	10.00	0.81	0.54
Convexity	1.1−38.72	12	32.43	9	24.32	97,260	10.00	3.24	2.43
	38.72−41.31	7	18.92	12	32.43	97,263	10.00	1.89	3.24
	41.31−43.19	11	29.73	10	27.03	97,261	10.00	2.97	2.70
	43.19−44.89	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	44.89−46.43	3	8.11	4	10.81	97,258	10.00	0.81	1.08
	46.43−47.93	0	0.00	0	0.00	97,263	10.00	0.00	0.00
	47.93−49.47	0	0.00	0	0.00	97,263	10.00	0.00	0.00
	49.47−51.24	1	2.70	0	0.00	97,262	10.00	0.27	0.00
	51.24−53.52	1	2.70	1	2.70	97,260	10.00	0.27	0.27
	53.52−79.32	1	2.70	0	0.00	97,263	10.00	0.27	0.00

Table A1. Cont.

Factor	Class	No. of SPC	% of SPC	No. of T	% of T	No. of Pixels in Domain	% of Pixels in Domain	Frequency Ratio of SPC	Frequency Ratio of T
Mass balance index (MBI)	−0.89−0.62	1	2.70	0	0.00	97,261	10.00	0.27	0.00
	−0.62−0.48	1	2.70	0	0.00	97,262	10.00	0.27	0.00
	−0.48−0.33	0	0.00	1	2.70	97,261	10.00	0.00	0.27
	−0.33−0.17	9	24.32	7	18.92	97,262	10.00	2.43	1.89
	−0.17−0.02	10	27.03	10	27.03	97,261	10.00	2.70	2.70
	−0.02−0.1	8	21.62	12	32.43	97,262	10.00	2.16	3.24
	0.1−0.33	3	8.11	3	8.11	97,261	10.00	0.81	0.81
	0.33−0.52	3	8.11	2	5.41	97,262	10.00	0.81	0.54
	0.52−0.68	1	2.70	0	0.00	97,261	10.00	0.27	0.00
0.68−1.09	1	2.70	2	5.41	97,262	10.00	0.27	0.54	
Slope angle (rad)	0−0.05	16	43.24	16	43.24	97,261	10.00	4.32	4.32
	0.05−0.12	8	21.62	11	29.73	97,262	10.00	2.16	2.97
	0.12−0.2	5	13.51	4	10.81	97,261	10.00	1.35	1.08
	0.2−0.26	3	8.11	3	8.11	97,262	10.00	0.81	0.81
	0.26−0.32	3	8.11	1	2.70	97,261	10.00	0.81	0.27
	0.32−0.37	0	0.00	0	0.00	97,262	10.00	0.00	0.00
	0.37−0.42	0	0.00	1	2.70	97,261	10.00	0.00	0.27
	0.42−0.48	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	0.48−0.56	1	2.70	0	0.00	97,261	10.00	0.27	0.00
0.56−1.05	0	0.00	0	0.00	97,262	10.00	0.00	0.00	
Slope height (m)	0.09−6.39	4	10.81	2	5.41	97,261	10.00	1.08	0.54
	6.39−8.36	13	35.14	13	35.14	97,262	10.00	3.51	3.51
	8.36−11.05	11	29.73	11	29.73	97,261	10.00	2.97	2.97
	11.05−14.61	4	10.81	6	16.22	97,262	10.00	1.08	1.62
	14.61−19.86	1	2.70	2	5.41	97,261	10.00	0.27	0.54
	19.86−26.58	3	8.11	2	5.41	97,262	10.00	0.81	0.54
	26.59−35.51	0	0.00	0	0.00	97,261	10.00	0.00	0.00
	35.51−48.23	0	0.00	1	2.70	97,262	10.00	0.00	0.27
	48.23−71.03	0	0.00	0	0.00	97,261	10.00	0.00	0.00
71.03−418.84	1	2.70	0	0.00	97,262	10.00	0.27	0.00	
Topographic texture	0.04−18.87	11	29.73	14	37.84	97,261	10.00	2.97	3.78
	18.87−29.08	11	29.73	15	40.54	97,259	10.00	2.97	4.05
	29.08−36.08	7	18.92	2	5.41	97,264	10.00	1.89	0.54
	36.08−41.43	4	10.81	3	8.11	97,261	10.00	1.08	0.81
	41.43−46.05	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	46.05−50.12	0	0.00	0	0.00	97,262	10.00	0.00	0.00
	50.12−53.76	1	2.70	0	0.00	97,261	10.00	0.27	0.00
	53.76−57.58	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	57.58−62.07	1	2.70	1	2.70	97,261	10.00	0.27	0.27
62.07−78.55	0	0.00	0	0.00	97,262	10.00	0.00	0.00	
Topographic position index (TPI)	−44.45−9.66	1	2.70	0	0.00	97,261	10.00	0.27	0.00
	−9.66−6	2	5.41	2	5.41	97,262	10.00	0.54	0.54
	−6−3.74	6	16.22	4	10.81	97,261	10.00	1.62	1.08
	−3.74−2.04	2	5.41	2	5.41	97,262	10.00	0.54	0.54
	−2.04−0.68	6	16.22	8	21.62	97,261	10.00	1.62	2.16
	−0.68−0.34	11	29.73	14	37.84	97,262	10.00	2.97	3.78
	0.34−2.78	4	10.81	4	10.81	97,261	10.00	1.08	1.08
	2.78−6.38	4	10.81	1	2.70	97,262	10.00	1.08	0.27
	6.38−11.46	0	0.00	1	2.70	97,261	10.00	0.00	0.27
11.46−73.57	1	2.70	1	2.70	97,262	10.00	0.27	0.27	
Topographic ruggedness index (TRI)	0−1.16	17	45.95	18	48.65	97,261	10.00	4.59	4.86
	1.16−3.02	7	18.92	10	27.03	97,262	10.00	1.89	2.70
	3.02−4.78	3	8.11	3	8.11	97,261	10.00	0.81	0.81
	4.78−6.14	6	16.22	3	8.11	97,262	10.00	1.62	0.81
	6.14−7.31	2	5.41	1	2.70	97,261	10.00	0.54	0.27
	7.31−8.45	0	0.00	0	0.00	97,262	10.00	0.00	0.00
	8.45−9.66	1	2.70	1	2.70	97,261	10.00	0.27	0.27
	9.66−11.1	1	2.70	0	0.00	97,261	10.00	0.27	0.00
	11.1−13.06	0	0.00	1	2.70	97,262	10.00	0.00	0.27
13.06−68.75	0	0.00	0	0.00	97,262	10.00	0.00	0.00	
Valley depth	0−9.29	2	5.41	1	2.70	97,261	10.00	0.54	0.27
	9.29−14.41	5	13.51	3	8.11	97,262	10.00	1.35	0.81
	14.41−20.19	0	0.00	2	5.41	97,261	10.00	0.00	0.54
	20.19−26.88	2	5.41	3	8.11	97,262	10.00	0.54	0.81
	26.88−34.55	5	13.51	7	18.92	97,261	10.00	1.35	1.89
	34.55−43.68	6	16.22	5	13.51	97,262	10.00	1.62	1.35
	43.68−54.86	3	8.11	6	16.22	97,261	10.00	0.81	1.62
	54.86−69.7	3	8.11	3	8.11	97,262	10.00	0.81	0.81
	69.7−93.65	4	10.81	1	2.70	97,261	10.00	1.08	0.27
93.65−351.98	6	15.79	6	16.22	97,262	10.00	1.89	1.62	

Table A1. Cont.

Factor	Class	No. of SPC	% of SPC	No. of T	% of T	No. of Pixels in Domain	% of Pixels in Domain	Frequency Ratio of SPC	Frequency Ratio of T
Flow path	0–60	12	32.43	12	32.43	93,864	9.65	3.36	3.36
	72.42–144.85	6	16.22	5	13.51	100,271	10.31	1.57	1.31
	150–229.7	7	18.92	8	21.62	92,423	9.50	1.99	2.28
	234.85–330	4	10.81	4	10.81	101,044	10.39	1.04	1.04
	332.13–434.55	3	8.11	4	10.81	95,153	9.78	0.83	1.11
	434.55–556.69	4	10.81	3	8.11	100,626	10.35	1.04	0.78
	556.69–704.55	1	2.70	1	2.70	96,887	9.96	0.27	0.27
	704.55–896.98	0	0.00	0	0.00	97,622	10.04	0.00	0.00
	896.98–1193.96	0	0.00	0	0.00	96,752	9.95	0.00	0.00
1193.97–3802.2	0	0.00	0	0.00	97,973	10.07	0.00	0.00	
LS factor	0–1.6	14	37.84	16	43.24	97,261	10.00	3.78	4.32
	1.6–4.63	10	27.03	11	29.73	97,262	10.00	2.70	2.97
	4.63–7.78	5	13.51	5	13.51	97,261	10.00	1.35	1.35
	7.78–10.72	4	10.81	2	5.41	97,262	10.00	1.08	0.54
	10.72–13.52	1	2.70	1	2.70	97,261	10.00	0.27	0.27
	13.52–16.34	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	16.34–19.42	1	2.70	0	0.00	97,261	10.00	0.27	0.00
	19.42–23.18	1	2.70	1	2.70	97,262	10.00	0.27	0.27
	23.18–29.11	0	0.00	0	0.00	97,261	10.00	0.00	0.00
29.11–304.73	0	0.00	0	0.00	97,262	10.00	0.00	0.00	
Forest type	Deciduous pine tree (PL)	3	8.11	3	8.11	240,329	24.71	0.33	0.33
	Pine forest (PK)	2	5.41	1	2.70	133,782	13.75	0.39	0.20
	Broadleaved forest (H)	0	0.00	0	0.00	181,244	18.63	0.00	0.00
	Mixed forest of soft and hardwood (M)	0	0.00	1	2.70	51,415	5.29	0.00	0.51
	Chestnut forest (Ca)	1	2.70	0	0.00	3168	0.33	8.30	0.00
	Non-forest (ND)	28	75.68	30	81.08	241,742	24.85	3.04	3.26
	Pine forest (D)	0	0.00	0	0.00	20,816	2.14	0.00	0.00
	Pinus rigida forest (PR)	3	8.11	2	5.41	75,341	7.75	1.05	0.70
	Farmland (L)	0	0.00	0	0.00	5653	0.58	0.00	0.00
	Needleleaf artificial forest (PD)	0	0.00	0	0.00	9602	0.99	0.00	0.00
	Left-over area (R)	0	0.00	0	0.00	6189	0.64	0.00	0.00
	Dentuded land (E)	0	0.00	0	0.00	119	0.01	0.00	0.00
	Broadleaved artificial forest (PH)	0	0.00	0	0.00	1664	0.17	0.00	0.00
	Poplar forest (Po)	0	0.00	0	0.00	244	0.03	0.00	0.00
	Grassland (LP)	0	0.00	0	0.00	877	0.09	0.00	0.00
	Oak forest (Q)	0	0.00	0	0.00	225	0.02	0.00	0.00
	Fine-grained wood (O)	0	0.00	0	0.00	31	0.00	0.00	0.00
Coniferous forest (C)	0	0.00	0	0.00	174	0.02	0.00	0.00	
Soil	Water	4	10.81	2	5.41	13,501	1.39	7.79	3.89
	Alluvium	7	18.92	8	21.62	97,462	10.02	1.89	2.16
	Regosol	2	5.41	2	5.41	76,862	7.90	0.68	0.68
	Red-yellow	4	10.81	6	16.22	107,414	11.04	0.98	1.47
	Lithosols	11	29.73	10	27.03	600,412	61.73	0.48	0.44
	Sierozem	7	18.92	7	18.92	60,429	6.21	3.05	3.05
	Planosol	1	2.70	1	2.70	245	0.03	107.29	107.29
	Other	1	2.70	1	2.70	16,290	1.67	1.61	1.61
Landcover	Urban	9	24.32	11	29.73	35,546	3.65	6.66	8.13
	Agriculture	15	40.54	14	37.84	179,578	18.46	2.20	2.05
	Forest	9	24.32	7	18.92	703,777	72.36	0.34	0.26
	Grass/Shrub	1	2.70	1	2.70	14,721	1.51	1.79	1.79
	Wetlands	1	2.70	1	2.70	4522	0.46	5.81	5.81
	Bare	0	0.00	0	0.00	13,032	1.34	0.00	0.00
	Water	2	5.41	3	8.11	21,465	2.21	2.45	3.67
Geology	Non	1	2.70	1	2.70	11,000	1.13	2.39	2.39
	Alluvium (Qa)	12	32.43	15	40.54	107,636	11.07	2.93	3.66
	Ganite-bearing granitic gneiss (PCEkgrtgn)	0	0.00	0	0.00	30,231	3.11	0.00	0.00
	Leucocratic gneiss (PCEklgn)	0	0.00	0	0.00	4314	0.44	0.00	0.00
	Migmatitic gneiss (PCEkmgn)	0	0.00	0	0.00	34,178	3.51	0.00	0.00
	Granite porphyry (Kgp)	0	0.00	0	0.00	9432	0.97	0.00	0.00
	Banded gneiss (PCEkbggn)	9	24.32	10	27.03	261,108	26.85	0.91	1.01
	Schists (PCEccs)	0	0.00	0	0.00	74,084	7.62	0.00	0.00
	Porphyroblastic gneiss (PCEpgn)	0	0.00	0	0.00	38,778	3.99	0.00	0.00
	Amphibolite (am)	0	0.00	0	0.00	1261	0.13	0.00	0.00
	Granite porphyry (Jgr)	14	37.84	10	27.03	253,162	26.03	1.45	1.04
Quartzite (Q)	0	0.00	0	0.00	25,512	2.62	0.00	0.00	

Table A1. Cont.

Factor	Class	No. of SPC	% of SPC	No. of T	% of T	No. of Pixels in Domain	% of Pixels in Domain	Frequency Ratio of SPC	Frequency Ratio of T
	Yangpyeong Igneous Complex (yic)	1	2.70	1	2.70	40,910	4.21	0.64	0.64
	Gneiss (PCEcgn)	0	0.00	0	0.00	73,955	7.60	0.00	0.00
	Diorite (Jdi)	0	0.00	0	0.00	3383	0.35	0.00	0.00
	Acidic (kad)	0	0.00	0	0.00	3671	0.38	0.00	0.00
	0–530.75	8	21.62	12	32.43	97,236	10.00	2.16	3.24
	531.6–1081.66	1	2.70	1	2.70	96,925	9.97	0.27	0.27
	1082.08–1611.36	1	2.70	3	8.11	97,419	10.02	0.27	0.81
	1612.2–2130.21	6	16.22	4	10.81	97,264	10.00	1.62	1.08
Distance from fault	2130.63–2673.2	3	8.11	4	10.81	97,368	10.01	0.81	1.08
	2674.21–3317.13	6	16.22	6	16.22	97,271	10.00	1.62	1.62
	3318.08–4440.4	3	8.11	0	0.00	97,341	10.01	0.81	0.00
	4440.91–7620.53	6	16.22	3	8.11	97,268	10.00	1.62	0.81
	7620.7–11187.71	2	5.41	2	5.41	97,259	10.00	0.54	0.54
	11187.91–17310.08	1	2.70	2	5.41	97,264	10.00	0.27	0.54

## References

- Oke, S.A.; Fourie, F. Guidelines to groundwater vulnerability mapping for Sub-Saharan Africa. *Groundw. Sustain. Dev.* **2017**, *5*, 168–177. [\[CrossRef\]](#)
- Noh, C. Average Savings of 339 Reservoirs in Gyeonggi-do. Available online: <http://www.todaykorea.co.kr/news/view.php?no=255887> (accessed on 8 September 2019).
- Naghibi, S.A.; Pourghasemi, H.R.; Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess.* **2016**, *188*, 44. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gaur, S.; Chahar, B.R.; Grailot, D. Combined use of groundwater modeling and potential zone analysis for management of groundwater. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 127–139. [\[CrossRef\]](#)
- Abdulkareem, J.H.; Pradhan, B.; Sulaiman, W.N.A.; Jamil, N.R. Quantification of runoff as influenced by morphometric characteristics in a rural complex catchment. *Earth Syst. Environ.* **2018**, *2*, 145–162. [\[CrossRef\]](#)
- Getirana, A.; Rodell, M.; Kumar, S.; Beaudoin, H.K.; Arsenault, K.; Zaitchik, B.; Save, H.; Bettadpur, S. GRACE improves seasonal groundwater forecast initialization over the US. *J. Hydrometeorol.* **2019**, *21*, 59–71. [\[CrossRef\]](#)
- Li, B.; Rodell, M.; Kumar, S.; Beaudoin, H.K.; Getirana, A.; Zaitchik, B.F.; de Goncalves, L.G.; Cossetin, C.; Bhanja, S.; Mukherjee, A. Global GRACE data assimilation for groundwater and drought monitoring: Advances and challenges. *Water Resour. Res.* **2019**, *55*, 7564–7586. [\[CrossRef\]](#)
- Nie, W.; Zaitchik, B.F.; Rodell, M.; Kumar, S.V.; Arsenault, K.R.; Li, B.; Getirana, A. Assimilating GRACE into a Land Surface Model in the presence of an irrigation-induced groundwater trend. *Water Resour. Res.* **2019**. [\[CrossRef\]](#)
- Becker, M.W. Potential for satellite remote sensing of ground water. *Groundwater* **2006**, *44*, 306–318. [\[CrossRef\]](#) [\[PubMed\]](#)
- Brunner, P.; Franssen, H.-J.H.; Kgotlhang, L.; Bauer-Gottwein, P.; Kinzelbach, W. How can remote sensing contribute in groundwater modeling? *Hydrogeol. J.* **2007**, *15*, 5–18. [\[CrossRef\]](#)
- Balbarini, N.; Bjerg, P.L.; Binning, P.J.; Christiansen, A.V. Modelling Tools for Integrating Geological, Geophysical and Contamination Data for Characterization of Groundwater Plumes. Ph.D. Thesis, Department of Environmental Engineering, Technical University of Denmark, Kgs., Lyngby, Denmark, 2017.
- Russoniello, C.; Michael, H.; Fernandez, C.; Andres, A.; He, C.; Madsen, J.A. *Investigation of Submarine Groundwater Discharge at Holts Landing State Park, Delaware: Hydrogeologic Framework, Groundwater Level and Salinity Observations*; Delaware Geological Survey, University of Delaware: Newark, DE, USA, 2017.
- Helaly, A.S. Assessment of groundwater potentiality using geophysical techniques in Wadi Allaqi basin, Eastern Desert, Egypt—Case study. *Nriag J. Astron. Geophys.* **2017**, *6*, 408–421. [\[CrossRef\]](#)
- Nampak, H.; Pradhan, B.; Manap, M.A. Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *J. Hydrol.* **2014**, *513*, 283–300. [\[CrossRef\]](#)

15. Ghorbani Nejad, S.; Falah, F.; Daneshfar, M.; Haghizadeh, A.; Rahmati, O. Delineation of groundwater potential zones using remote sensing and GIS-based data-driven models. *Geocarto Int.* **2017**, *32*, 167–187. [[CrossRef](#)]
16. Sameen, M.I.; Pradhan, B.; Lee, S. Self-learning random forests model for mapping groundwater yield in data-scarce areas. *Nat. Resour. Res.* **2019**, *28*, 757–775. [[CrossRef](#)]
17. Elfadaly, A.; Attia, W.; Lasaponara, R. Monitoring the environmental risks around Medinet Habu and Ramesseum Temple at West Luxor, Egypt, using remote sensing and GIS techniques. *J. Archaeol. Method Theory* **2018**, *25*, 587–610. [[CrossRef](#)]
18. Elmahdy, S.I.; Mohamed, M.M. Automatic detection of near surface geological and hydrological features and investigating their influence on groundwater accumulation and salinity in southwest Egypt using remote sensing and GIS. *Geocarto Int.* **2015**, *30*, 132–144. [[CrossRef](#)]
19. Fernandez, P.; Delgado, E.; Lopez-Alonso, M.; Poyatos, J.M. GIS environmental information analysis of the Darro River basin as the key for the management and hydrological forest restoration. *Sci. Total Environ.* **2018**, *613*, 1154–1164. [[CrossRef](#)]
20. Lee, J. Review of remote sensing studies on groundwater resources. *Korean J. Remote Sens.* **2017**, *33*, 855–866.
21. Hadžić, E.; Lazović, N.; Mulaomerović-Šeta, A. Application of mathematical models in defining optimal groundwater yield. *Procedia Environ. Sci.* **2015**, *25*, 112–119. [[CrossRef](#)]
22. Golkarian, A.; Naghibi, S.A.; Kalantar, B.; Pradhan, B. Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. *Environ. Monit. Assess.* **2018**, *190*, 149. [[CrossRef](#)]
23. Kim, J.-C.; Jung, H.-S.; Lee, S. Groundwater productivity potential mapping using frequency ratio and evidential belief function and artificial neural network models: Focus on topographic factors. *J. Hydroinformatics* **2018**, *20*, 1436–1451. [[CrossRef](#)]
24. Rahmati, O.; Naghibi, S.A.; Shahabi, H.; Bui, D.T.; Pradhan, B.; Azareh, A.; Rafiei-Sardooi, E.; Samani, A.N.; Melesse, A.M. Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches. *J. Hydrol.* **2018**, *565*, 248–261. [[CrossRef](#)]
25. Kim, J.-C.; Jung, H.-S.; Lee, S. Spatial mapping of the groundwater potential of the geum river basin using ensemble models based on remote sensing images. *Remote Sens.* **2019**, *11*, 2285. [[CrossRef](#)]
26. Lee, S.; Hyun, Y.; Lee, M.-J. Groundwater potential mapping using data mining models of big data analysis in Goyang-si, South Korea. *Sustainability* **2019**, *11*, 1678. [[CrossRef](#)]
27. Lee, S.; Lee, C.-W.; Kim, J.-C. Groundwater productivity potential mapping using logistic regression and boosted tree models: The case of Okcheon city in Korea. In *Advances in Remote Sensing and Geo Informatics Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 305–307.
28. Das, S. Comparison among influencing factor, frequency ratio, and analytical hierarchy process techniques for groundwater potential zonation in Vaitarna basin, Maharashtra, India. *Groundw. Sustain. Dev.* **2019**, *8*, 617–629. [[CrossRef](#)]
29. Trabelsi, F.; Lee, S.; Khlifi, S.; Arfaoui, A. Frequency ratio model for mapping groundwater potential zones using gis and remote sensing; Medjerda watershed Tunisia. In *Advances in Sustainable and Environmental Hydrology, Hydrogeology, Hydrochemistry and Water Resources*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 341–345.
30. Arulbalaji, P.; Padmalal, D.; Sreelash, K. GIS and AHp techniques based delineation of groundwater potential zones: A case study from southern Western Ghats, India. *Sci. Rep.* **2019**, *9*, 2082. [[CrossRef](#)] [[PubMed](#)]
31. Nami, M.; Jaafari, A.; Fallah, M.; Nabiuni, S. Spatial prediction of wildfire probability in the Hyrcanian ecoregion using evidential belief function model and GIS. *Int. J. Environ. Sci. Technol.* **2018**, *15*, 373–384. [[CrossRef](#)]
32. Chen, W.; Li, H.; Hou, E.; Wang, S.; Wang, G.; Panahi, M.; Li, T.; Peng, T.; Guo, C.; Niu, C. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Sci. Total Environ.* **2018**, *634*, 853–867. [[CrossRef](#)] [[PubMed](#)]
33. Naghibi, S.A.; Pourghasemi, H.R.; Abbaspour, K. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theor. Appl. Climatol.* **2018**, *131*, 967–984. [[CrossRef](#)]

34. Lee, S.; Hong, S.-M.; Jung, H.-S. GIS-based groundwater potential mapping using artificial neural network and support vector machine models: The case of Boryeong city in Korea. *Geocarto Int.* **2018**, *33*, 847–861. [[CrossRef](#)]
35. Josephs-Afoko, D.; Godfrey, S.; Campos, L.C. Assessing the performance and robustness of the UNICEF model for groundwater exploration in Ethiopia through application of the analytic hierarchy process, logistic regression and artificial neural networks. *Water Sa* **2018**, *44*, 365–376. [[CrossRef](#)]
36. Kumar, A.; Krishna, A.P. Assessment of groundwater potential zones in coal mining impacted hard-rock terrain of India by integrating geospatial and analytic hierarchy process (AHP) approach. *Geocarto Int.* **2018**, *33*, 105–129. [[CrossRef](#)]
37. Kordestani, M.D.; Naghibi, S.A.; Hashemi, H.; Ahmadi, K.; Kalantar, B.; Pradhan, B. Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeol. J.* **2019**, *27*, 211–224. [[CrossRef](#)]
38. Miraki, S.; Zanganeh, S.H.; Chapi, K.; Singh, V.P.; Shirzadi, A.; Shahabi, H.; Pham, B.T. Mapping groundwater potential using a novel hybrid intelligence approach. *Water Resour. Manag.* **2019**, *33*, 281–302. [[CrossRef](#)]
39. Pham, B.T.; Jaafari, A.; Prakash, I.; Singh, S.K.; Quoc, N.K.; Bui, D.T. Hybrid computational intelligence models for groundwater potential mapping. *Catena* **2019**, *182*, 104101. [[CrossRef](#)]
40. Lee, Y.-S.; Park, S.-H.; Jung, H.-S.; Baek, W.-K. Classification of Natural and Artificial Forests from KOMPSAT-3/3A/5 Images Using Artificial Neural Network. *Korean J. Remote Sens.* **2018**, *34*, 1399–1414.
41. Bauer, E.; Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.* **1999**, *36*, 105–139. [[CrossRef](#)]
42. Gyeonggi Research Institute. *Improvements of the Groundwater Management System in Gyeonggi-do. 2018*; Gyeonggi Research Institute: Gyeonggi-do, South Korea, 2018.
43. K Water. *Groundwater Annual Report. 2017*; Ministry of Environment, Water Policy Coordination Division: Sejong-si, South Korea, 2017.
44. Fallon, A.; Villholth, K.; Conway, D.; Lankford, B.; Ebrahim, G. Agricultural groundwater management strategies and seasonal climate forecasting: Perceptions from Mogwadi (Dendron), Limpopo, South Africa. *J. Water Clim. Chang.* **2019**, *10*, 142–157. [[CrossRef](#)]
45. K Water. *Groundwater Basic Survey Report-Yangpyeong-Gun*; Ministry of Environment, Water Policy Coordination Division: Sejong-si, South Korea, 2008.
46. Miralles-Wilhelm, F.; Hejazi, M.; Kim, S.; Yonkofski, C.; Watson, D.; Kyle, P.; Liu, Y.; Vernon, C.; Delgado, A.; Edmonds, J. *Water for Food and Energy Security: An Assessment of the Impacts of Water Scarcity on Agricultural Production and Electricity Generation in the Middle East and North Africa*; World Bank: Washington, DC, USA, 2018.
47. Kalantar, B.; Al-Najjar, H.A.; Pradhan, B.; Saeidi, V.; Halin, A.A.; Ueda, N.; Naghibi, S.A. Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water* **2019**, *11*, 1909. [[CrossRef](#)]
48. Mogaji, K.A.; Lim, H.S. Development of groundwater favourability map using GIS-based driven data mining models: An approach for effective groundwater resource management. *Geocarto Int.* **2018**, *33*, 397–422. [[CrossRef](#)]
49. Naghibi, S.A.; Pourghasemi, H.R. A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water Resour. Manag.* **2015**, *29*, 5217–5236. [[CrossRef](#)]
50. SAGA-GIS System for Automated Geoscientific Analyses. Available online: [www.sagagis.org](http://www.sagagis.org) (accessed on 9 January 2020).
51. Al-Abadi, A.M.; Pradhan, B.; Shahid, S. Prediction of groundwater flowing well zone at An-Najif Province, central Iraq using evidential belief functions model and GIS. *Environ. Monit. Assess.* **2016**, *188*, 549. [[CrossRef](#)] [[PubMed](#)]
52. Lee, S.; Lee, C.-W. Application of decision-tree model to groundwater productivity-potential mapping. *Sustainability* **2015**, *7*, 13416–13432. [[CrossRef](#)]
53. Lee, S.; Lee, S.; Lee, M.-J.; Jung, H.-S. Spatial assessment of urban flood susceptibility using data mining and geographic information System (GIS) tools. *Sustainability* **2018**, *10*, 648. [[CrossRef](#)]
54. Lee, S.; Pradhan, B. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* **2007**, *4*, 33–41. [[CrossRef](#)]

55. Pradhan, B. Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. *J. Indian Soc. Remote Sens.* **2010**, *38*, 301–320. [[CrossRef](#)]
56. Sujatha, E.R.; Rajamanickam, V.; Kumaravel, P.; Saranathan, E. Landslide susceptibility analysis using probabilistic likelihood ratio model—A geospatial-based study. *Arab. J. Geosci.* **2013**, *6*, 429–440. [[CrossRef](#)]
57. Lee, S.; Lee, M.-J.; Lee, S. Spatial prediction of urban landslide susceptibility based on topographic factors using boosted trees. *Environ. Earth Sci.* **2018**, *77*, 656. [[CrossRef](#)]
58. Bresfelean, V.P. Analysis and predictions on students' behavior using decision trees in Weka environment. In Proceedings of the 29th International Conference on Information Technology Interfaces, Cavtat, Croatia, 25–28 June 2007; pp. 25–28.
59. Zhao, Y.; Zhang, Y. Comparison of decision tree methods for finding active objects. *Adv. Space Res.* **2008**, *41*, 1955–1959. [[CrossRef](#)]
60. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and regression trees. *Wadsworth Int. Group* **1984**, *37*, 237–251.
61. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *J. R. Stat. Soc.* **1980**, *29*, 119–127. [[CrossRef](#)]
62. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [[CrossRef](#)]
63. Quinlan, J.R. *C4. 5: Programs for machine learning*; Elsevier: Amsterdam, The Netherlands, 2014.
64. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [[CrossRef](#)]
65. Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* **1995**, *27*, 326–327. [[CrossRef](#)]
66. Schaffer, C. Overfitting avoidance as bias. *Mach. Learn.* **1993**, *10*, 153–178. [[CrossRef](#)]
67. Altman, D.G.; Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* **1994**, *308*, 1552. [[CrossRef](#)] [[PubMed](#)]
68. Lee, S.; Lee, M.-J.; Jung, H.-S. Data mining approaches for landslide susceptibility mapping in Umyeonsan, Seoul, South Korea. *Appl. Sci.* **2017**, *7*, 683. [[CrossRef](#)]
69. Hsu, F.-M.; Lin, Y.-T.; Ho, T.-K. Design and implementation of an intelligent recommendation system for tourist attractions: The integration of EBM model, Bayesian network and Google Maps. *Expert Syst. Appl.* **2012**, *39*, 3257–3264. [[CrossRef](#)]
70. Kuncheva, L.I. Classifier ensembles for changing environments. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 9–11 June 2004; pp. 1–15.
71. Bui, D.T.; Ho, T.-C.; Pradhan, B.; Pham, B.-T.; Nhu, V.-H.; Revhaug, I. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environ. Earth Sci.* **2016**, *75*, 1101.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).