IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Groundwater Potential Zone Assessment Using Gis And Ensemble Machine Learning Models In Parts Of The Mahanadi River Basin, Sambalpur, Odisha

Debasis Sahoo[1], *Jagadish Kumar Tripathy[1], Priyanka Sahu[1], Manas Ranjan Jena[1], Sunanda Biswal[1]

[1]Department of Earth Sciences, Sambalpur University, Jyoti Vihar, Burla, Sambalpur, Odisha-768019, India

Abstract: Groundwater forms an important freshwater commodity in the semi-arid and hard-rock provinces of India, with its distribution governed by complex relationships between geological, geomorphological, topographical and climatic factors. The present investigation outlines a comprehensive framework for spatially prognosticating groundwater potential zones (GWPZ) in a selected section of the Mahanadi River Basin, Sambalpur district, Odisha, using GIS-based thematic analysis in conjunction with comparative machine learning methodologies. Twelve conditioning variables relevant to groundwater availability, including geology, geomorphology, land use/land cover, soil characteristics, elevation, slope, curvature, topographic position index (TPI), topographic wetness index (TWI), lineament density, drainage density and precipitation, were processed and analysed as part of a GIS environment. The identification of groundwater potential was completed through the Analytical Hierarchy Process (AHP) using the combination of seven predictive modelling paradigms: Frequency Ratio (FR), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and XGBoost. Model efficacy was evaluated based on receiver operating characteristic-area under the curve (ROC-AUC), ground truth validation accuracy and Cohen's Kappa statistics. Findings show that the Artificial Neural Network (ANN; Accuracy= 93.75% and Kappa= 0.90) and Random Forest (RF; Accuracy= 91.25% and Kappa= 0.86) models outperform the other approaches in groundwater potential forecast, whereas the Frequency Ratio (FR), although it shows a discernible trade-off of the AUC, has lost its spatial reliability. Good groundwater potential zones are mostly coincident with landscapes that have low slope, high lineament density, good geomorphologic characteristics, and moderate drainage density. These results demonstrate that the combination of GIS and machine learning models significantly enhances the accuracy and reliability of groundwater potential mapping and therefore provides an effective decision support tool for sustainable groundwater management and site selection for groundwater extraction systems in river basin areas.

Index Terms **-** Groundwater potential, GIS, Machine learning, Random Forest, ANN, ROC–AUC, Mahanadi River Basin.

INTRODUCTION

Groundwater is a vital freshwater resource for sustaining domestic, agricultural and industrial demands for water especially in semi-arid and monsoon-dependent regions of India (Sishodia et al., 2018). In the river basin systems like the Mahanadi River Basin, groundwater availability is a key factor in counteracting the seasonal variation in surface water availability (Kumar & Bassi, 2021). However, the rapid population growth, intensification of agriculture, climatic variations and unplanned groundwater abstracting have exerted

collective pressure on water resources in the subsurface, which calls for scientifically sound approaches for groundwater assessment and management (Scanlon et al., 2023). Among the various districts in the basin, Sambalpur is a hydrogeologically complex and socio-economically significant area where the spatial variation of lithology, geomorphology, and land use has a great control over groundwater occurrence and productivity. Traditionally, groundwater potential zone (GWPZ) mapping has been based on traditional hydrogeologists survey and multi-criteria decision-making (MCDM) methods combined in Geographic Information Systems (GIS) (Pandey et al., 2022). Techniques like weighted overlay analysis, analytic hierarchy process (AHP) and frequency ratio models have been widely used to delineate zones of groundwater potential using thematic layers such as geology, geomorphology, slope, drainage density and land use/land cover. While these methods have provided some valuable insights, they are often limited by subjectivity in weight assignment, linear assumptions, and limited ability to capture complex and non-linear interactions between controlling factors. As a result, their predictive ability and their ability to generalize across heterogeneous terrains is limited (Vayadande et al., 2024). Supervised ML models like Random Forest, Support Vector Machine, Gradient Boosting, Artificial Neural Networks, Logistic Regression, etc., have shown excellent performance in dealing with high-dimensional data, complex non-linear relationships and interactions among the multiple hydrogeological and topographic variables (Jari et al., 2023). These data-driven approaches enable the possibility of learning in an objective way from observed groundwater occurrence data (e.g. well yield or groundwater depth), and thus reducing the reliance on expert judgement alone (Roy et al., 2023). Despite the increasing global applications, multiple ML models have been comparatively evaluated using comprehensive thematic datasets for river basin applications, but not for Indian river basins, especially in eastern India and the Mahanadi River Basin (Raj & Gopikrishnan, 2024). Moreover, extant studies in the Sambalpur region and adjoining parts of Odisha are often localized, method-specific or based on a limited number of thematic layers. There is a notable research gap in the systematic assessment of the relative importance and predictive capability of an integrated set of geomorphic, hydrological, topographic and climatic variables (e.g. curvature, topographic position index (TPI), topographic wetness index (TWI), lineament density and rainfall) in addition to the conventional factors. Addressing this gap is critical for the accuracy of groundwater potential mapping and groundwater management and planning at the basin scale. Against this background, the present study is an attempt to contribute to the spatial prediction of groundwater potential in parts of the Mahanadi River Basin in Sambalpur district, Odisha using a comparative machine learning framework coupled with GIS. The study makes use of twelve thematic layers, i.e., geology, geomorphology, land use/land cover, soil, elevation, slope, curvature, TPI, TWI, lineament density, drainage density and rainfall to account for the multifaceted controls on groundwater occurrence. The main research questions that will guide this study are: (i) to what extent can various machine learning algorithms be used for predicting zones of groundwater potential using integrated GIS-based thematic data; (ii) which models perform better in terms of predictive performance and robustness in the hydrogeological scenario of Mahanadi River Basin; and (iii) which conditioning factors have the greatest influence on groundwater potential in the study area. The rationale of this research is that it can lay the foundation for a scientifically sound and data rich methodology of groundwater potential assessment that can be used for sustainable water resource management, well siting, and policy formulation in data-poor areas. By the comparative assessment of various ML models, along with quantification of factor importance, the study adds methodological knowledge to the growing body of literature on groundwater modeling and also provides region-wise knowledge that is relevant to the eastern Indian hard rock and alluvial environments (Malakar et al., 2020). Methodologically, the study combines remote sensing-derived and ancillary spatial datasets in a GIS environment, applies various supervised machine learning algorithms for groundwater potential prediction and assesses model performance using statistical validation measures. The resulting groundwater potential maps are divided into three distinct potential zones to make it easier to interpret and put to practical use.

STUDY AREA

This study focuses on the western part of Sambalpur district Odisha, India, that covers an area of around 308.27 km^2. Geographically, the study area lies between latitudes 21.38 ° to 21.61 ° N and longitudes 83.93° to 84.15° E and comes under toposheets F44R14, F44R15, F45M2 and F45M3 of the Survey of India (Fig. 1). Belonging to the Mahanadi River Basin, one of the major east-flowing river system of peninsular India, the region presents a heterogeneous array of hydrogeological and geomorphological characteristics that make it an appropriate subject for groundwater potential assessment study using GIS and machine learning techniques. Physiographically in character, the land is characterized by undulating to moderately dissected topography, with elevation gradually decreasing to the Mahanadi River and its tributaries. Drainage plays out

mostly in dendritic to subdendritic patterns, suggestive of structural control imposed by underlying lithology and fracture structures. Spatial variation in drainage density affects runoff, infiltration and groundwater recharge mechanisms. The area has a tropical monsoon climate with most of the precipitation occurring in the southwest monsoon (June - September). Average annual rainfall is 1400 - 1600 mm, being a very important parameter in groundwater recharge dynamics (CGWB, 2023). Geologically, the study area is dominated by Precambrian crystalline complexes principally that of granitic and gneissic formations interspersed with localised alluvial deposits juxtaposed to riverine corridors. These hard rock aquifers are characterised by secondary porosity, where groundwater occurrence is controlled by the degree of weathering, fracturing and jointing. The weathered and fractured zones form the main hydrogeological active zones and the well yields show large variability depending on the structural density and the geomorphological context (CGWB, 2024). Prior hydrogeological evaluation of the Sambalpur region has highlighted the significant role of lineaments and geomorphological units comprising pediplains, valley fills, and buried pediments, on potential groundwater availability in the region (Singandhupe & Sethi, 2016). The edaphic profile is dominated and characterized by red and lateritic soils interspersed with localized occurrence of alluvial deposits near riverine areas. These soils show moderate to low permeability values and therefore affect infiltration dynamics and recharge rates. Land-use/land-cover (LULC) regimes are mainly of an agrarian nature containing croplands, fallow fields, patches of forests, and gradually expanding built-up areas. Alterations and changes in LULC, specifically intensification of agriculture and urban sprawl have recorded as affecting groundwater extraction and recharge conditions along sections of the Mahanadi Basin (Jha et al. 2010). In terms of groundwater management, Sambalpur district falls under the categories of safe to semi-critical with respect to groundwater development; however, some stress zones have been identified on a local basis, which has been attributed to escalating demands for irrigation coupled with the seasonal variation in replenishment to groundwater. (CGWB, 2024). The large-scale spatial heterogeneity of lithology, topography, drainage and land use makes a conventional groundwater assessment difficult and hence highlights the need for sophisticated modeling techniques, which require data-driven modeling. Contemporary investigations in river basins groundwater evaluation related to the worth of gathering topographic indices (slope, curvature, Topographic Wetness Index-TWI and Topographic Position Index (TPI))-downside along with structural (lineament density) and climatic parameters (precipitation) within a GIS-based machine learning framework for higher predictive power (Naghibi et al., 2017; Rahmati et al., 2019). The selected study area, with its complex hard rock hydrogeology and monsoon-based recharge regime acts as a suitable testing ground for evaluating the comparative performance of machine learning models used for groundwater potential mapping. Overall, the summation of the hydrogeological complex, socioeconomic dependence of groundwater and spatial heterogeneities of controlling determinants in western Sambalpur district makes the present study even more significant in scientific and practical applications. The creation of dependable groundwater potential maps using multi-source geospatial datasets through machine learning methodologies is expected to aid in the sustainable groundwater governance, best practices for well-siting and basin-scale water resources management in the Mahanadi River Basin.

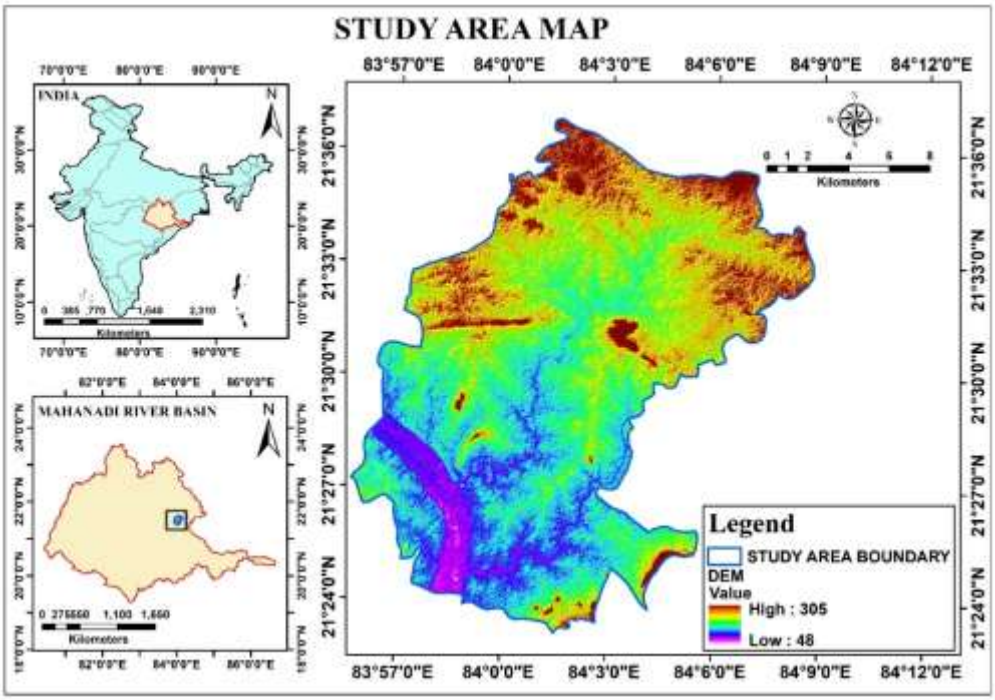

Fig. 1 Study area map.

METHODOLOGY

An integrated geographical information system (GIS) and Machine Learning (ML) approach was applied for the spatial prediction of groundwater potential zones (GWPZ) in selected sub-areas of the Mahanadi River Basin, in Sambalpur district in the state of Odisha. The methodological workflow (Fig. 2) includes data acquisition, thematic layers preparation, knowledge-driven groundwater potential mapping through Analytical Hierarchy Process (AHP), data-driven modelling using different machine learning algorithms, model validation, uncertainty assessment and ensemble fusion for the generation of the final groundwater potential map. Comparable integrated frameworks have been widely recommended for groundwater potential mapping from complex hydrogeological scenarios (Naghibi, et al., 2017; Rahmati, et al., 2019). All the spatial data processing and thematic layer generation have been performed with ArcGIS Desktop 10.8 (Esri, 2020). DEM-based terrain and hydrological indices including slope and curvature, Topographic Position Index (TPI), Topographic Wetness Index (TWI) and drainage density were extracted in Spatial Analyst tools and Hydrology Tools. Lineament density was calculated using kernel density estimation method while drainage networks were delineated using standard flow direction and flow accumulation procedures. Machine-learning modelling and statistical analysis were implemented with Python 3.11 by mainly using open-source tools such as NumPy, Pandas, scikit-learn, XGBoost and TensorFlow/Keras. Hyperparameter tuning and cross-validation were implemented using the model selection utilities of scikit-learn. The combination of predictors generated from GIS information and ML-based algorithms implemented in Python ensured reproducibility, scalability and robustness of methodology (Pedregosa et al., 2011). Twelve groundwater conditioning factors (Fig. 3), including geology, geomorphology, land use/land cover (LULC), soil, elevation, slope, curvature, TPI, TWI, lineament density (LD), drainage density (DD) and rainfall were chosen based on hydrogeological relevance and supporting studies (Machiwal et al., 2011; Prasad et al., 2020). In order to enable subsequent modelling, each thematic layer was classified into groundwater favourability subclasses.

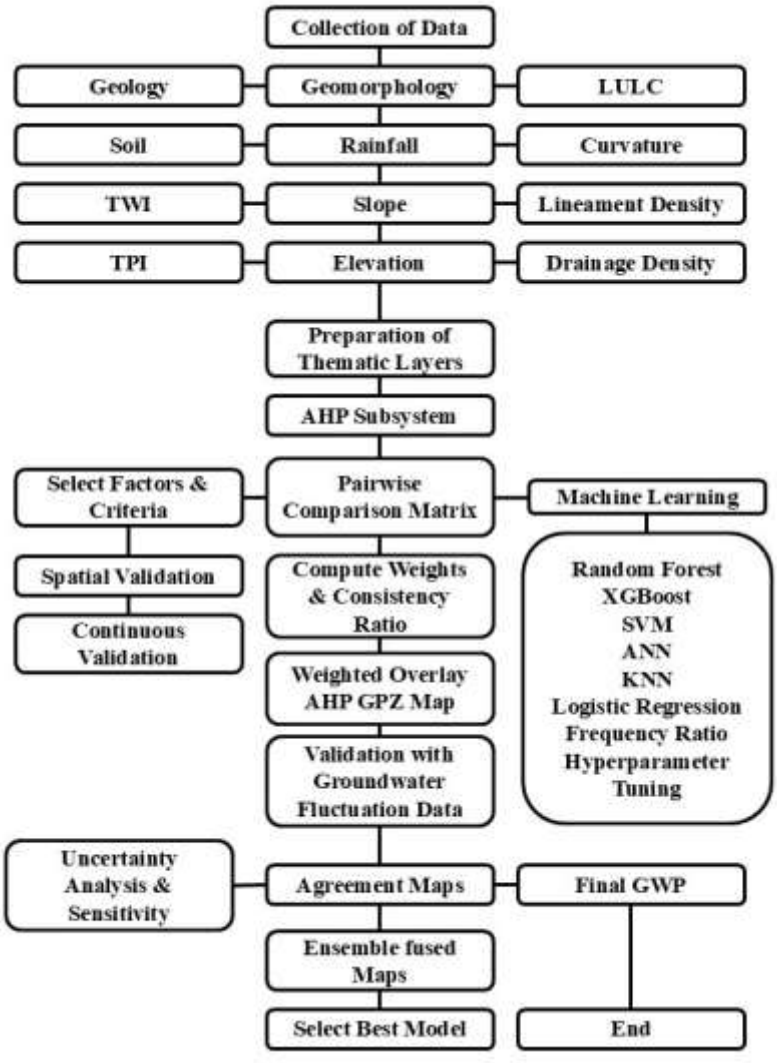

Fig. 2 Methodological workflow.

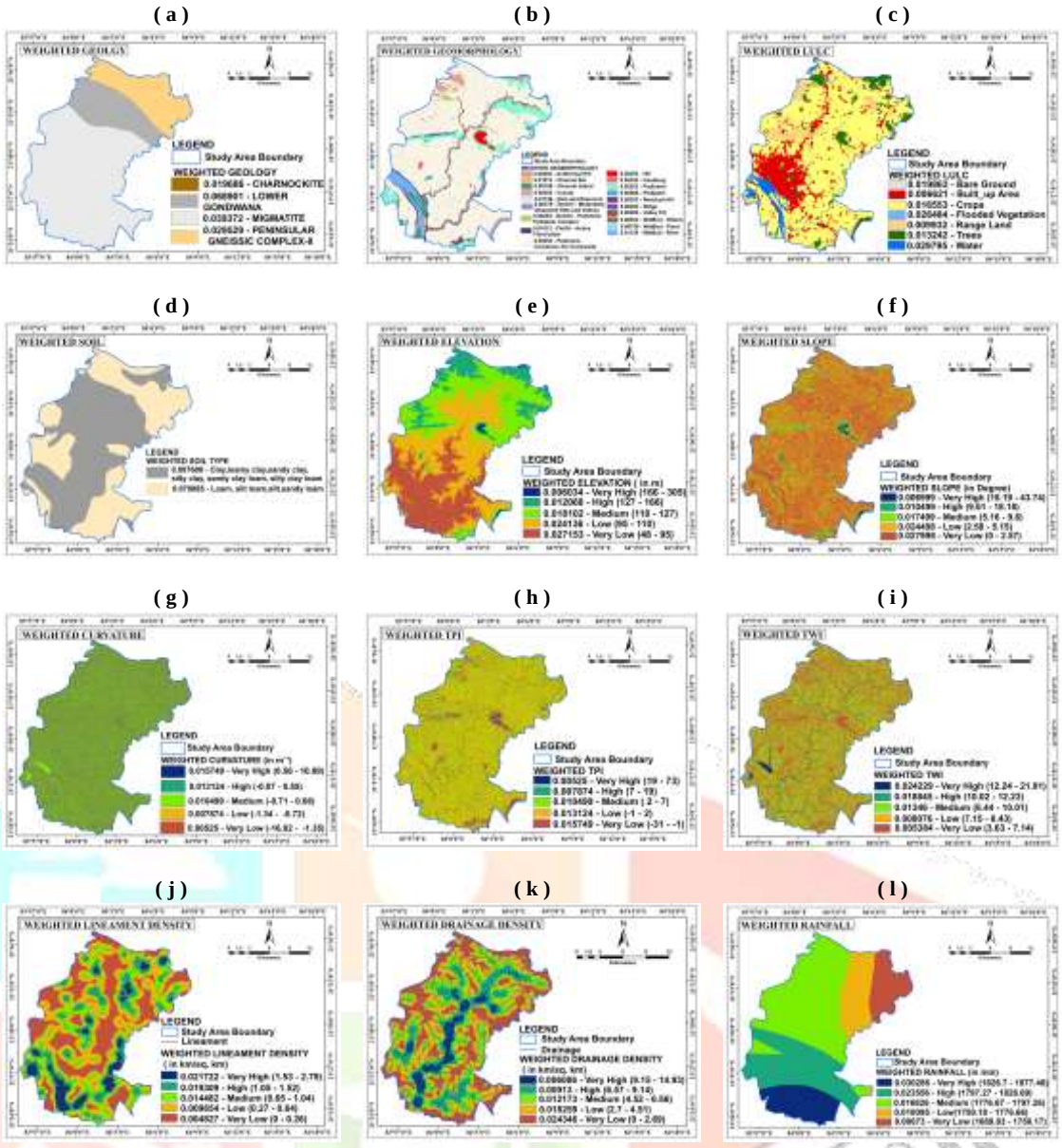

Fig. 3. Figure showing thematic layers (a: geology, b: geomorphology, c: LULC, d: soil, e: elevation, f: slope, g: curvature, h: TPI, i: TWI, j: lineament density, k: drainage density, l: rainfall.

Analytical Hierarchy Process (AHP)

Pairwise Comparison Matrix and Weight Derivation

An Analytic Hierarchy Process (AHP) was used to determine the relative importance of the groundwater conditioning factors by pairwise comparison matrix (Table 1) in accordance with Saaty (1980). The resulting normalized principal eigenvector of the comparison matrix is used as the weight vector of the thematic layers.

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ \frac{1}{a_{12}} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{1n}} & \frac{1}{a_{2n}} & \cdots & 1 \end{bmatrix}$$

Table 1 Pairwise comparison matrix

Factors	Assigned Weight	Slope	Geology	Geomorphology	LULC	Rainfall	Soil	TWI	TPI	LD	DD	Elevation	Curvature	Geometric Mean	Actual Weight	Wt %
Slope	5	1	0.56	0.83	0.71	1.00	1.00	1.25	1.67	1.25	1.25	1.00	1.67	1.05	0.09	8
Geology	9	1.80	1	1.50	1.29	1.80	1.80	2.25	3.00	2.25	2.25	1.80	3.00	1.89	0.16	15
Geomorphology	6	1.20	0.67	1	0.86	1.20	1.20	1.50	2.00	1.50	1.50	1.20	2.00	1.26	0.10	10
LULC	7	1.40	0.78	1.17	1	1.40	1.40	1.75	2.33	1.75	1.75	1.40	2.33	1.47	0.12	12
Rainfall	5	1.00	0.56	0.83	0.71	1	1.00	1.25	1.67	1.25	1.25	1.00	1.67	1.05	0.09	8
Soil	5	1.00	0.56	0.83	0.71	1.00	1	1.25	1.67	1.25	1.25	1.00	1.67	1.05	0.09	8
TWI	4	0.80	0.44	0.67	0.57	0.80	0.80	1	1.33	1.00	1.00	0.80	1.33	0.84	0.07	7
TPI	3	0.60	0.33	0.50	0.43	0.60	0.60	0.75	1	0.75	0.75	0.60	1.00	0.63	0.05	5
LD	4	0.80	0.44	0.67	0.57	0.80	0.80	1.00	1.33	1	1.00	0.80	1.33	0.84	0.07	7
DD	4	0.80	0.44	0.67	0.57	0.80	0.80	1.00	1.33	1.00	1	0.80	1.33	0.84	0.07	7

Elevation	5	1.00	0.56	0.83	0.71	1.00	1.00	1.25	1.67	1.25	1.25	1	1.67	1.05	0.09	8
Curvature	3	0.60	0.33	0.50	0.43	0.60	0.60	0.75	1.00	0.75	0.75	0.60	1	0.63	0.05	5

Consistency Evaluation

Reliability of judgments was assessed in the form of the Consistency Index (CI) and Consistency Ratio (CR) proposed by Saaty (1980).

$$CI = \frac{\lambda_{\max} - n}{n - 1} \quad (1)$$

$$CR = \frac{CI}{RI} \quad (2)$$

A CR value ≤ 0.1 was considered acceptable.

AHP-Based Groundwater Potential Index

The substantial groundwater potential index (GWPI) was calculated using a weighted linear combination of relevant hydrological variables with each being given a proportional weight (Table 2) in accordance with the relative impact of the variables given.

$$GWPI = \sum_{i=1}^{n} w_i \times r_i \quad (3)$$

where w_i is the weight of factor i and r_i is the rank of the corresponding subclass (Adiat et al., 2012).

Table 2 Rank and weights of thematic layers

Factors	Weight	Rank	Factors	Weight	Rank
Geology			**Soil**		
Peninsular gneissic complex-II	15	3	Sandy loam, loam, silt loam, silt	08	7
Migmatite		4	Clay, loamy clay, sandy clay, silty clay, sandy clay loam, silty clay loam		4
Charnockite		2	**Slope**		
Lower Gondwana		7	Very Low (0-2.57)	08	8
Geomorphology			Low (2.58-5.15)		7
Pediment-Corestone-Tor Composite	10	2	Medium (5.16-9.6)		5
WatBod – Pond		6	High (9.61-18.18)		3
Residual Hill		2	Very High (18.19-43.74)		2
Pediment		2	**Rainfall**		
Antiformal Hill		2	Very Low (1689.83-1750.17)	08	2
Cuesta		2	Low (1750.18-1776.66)		3
DenOri - Pediment-Pediplain Complex		2	Medium (1776.67-1797.26)		5
Valley Fill		7	High (1797.27-1826.69)		7
Inselberg		2	Very High (1826.7-1877.46)		9
Hill		2	**TWI**		
Pediplain		4	Very Low (3.63-7.14)	07	2
DenOri - Moderately Dissected Hills and Valleys		3	Low (7.15-8.43)		3
FluOri - Active Flood plain		8	Medium (8.44-10.01)		5
WatBod – River		9	High (10.02-12.23)		7
Ridge		2	Very High (12.24-21.91)		9

Channel Island		6
Dam and Reservoir		9
WatBod – Others		5
Channel Bar		8
LULC		
Water	12	9
Trees		4
Flooded Vegetation		8
Crops		5
Built-up Area		2
Bare Ground		6
Range Land		3
LD		
Very Low (0 - 0.26)	07	2
Low (0.27- 0.64)		4
Medium (0.65- 1.04)		6
High (1.05- 1.52)		8
Very High (1.53- 2.79)		9
DD		
Very Low (0 – 2.69)	07	8
Low (2.7 – 4.51)		6
Medium (4.52 – 6.56)		4
High (6.57 – 9.14)		3
Very High (9.15 – 14.93)		2

TPI		
Very Low (-31- -1)	05	6
Low (-1-2)		5
Medium (2-7)		4
High (7-19)		3
Very High (19-73)		2
Curvature		
Very Low (-16.02- -1.35)	05	2
Low (-1.34- -0.72)		3
Medium (-0.71- -0.08)		4
High (-0.07-0.55)		5
Very High (0.56-10.89)		6
Elevation		
Very Low (48- 95)	08	9
Low (95- 110)		8
Medium (110- 127)		6
High (127-166)		4
Very High (166-305)		2

Groundwater Inventory and L

Groundwater inventory information, including location of wells and information on groundwater fluctuation and yield, was obtained from the CGWB records and some in field observations. The total 80 well inventory data were divided into training data and validation data. Binary or continuous labels for the occurrence of groundwater were then derived to serve as the dependent feature of machine learning models in accordance with the established practices in data-driven groundwater potential mapping (Naghibi et al., 2017).

Machine Learning Models

Frequency Ratio (FR)

The Frequency Ratio methodology is used to offer a quantitative measurement of the spatial association of groundwater occurrence in relation to pertinent conditioning factors (Bonham-Carter, 1994):

$$FR = \frac{\left(\frac{N_{gw,i}}{N_{gw}}\right)}{\left(\frac{N_i}{N}\right)} \quad (4)$$

where $N_{gw,i}$is the number of groundwater occurrence pixels in class i, $N_{gw}$is the total groundwater pixels, $N_i$is the number of pixels in class i, and $N$is the total number of pixels. The groundwater potential index is obtained by summing FR values across all factors.

The cumulative FR index was used to generate groundwater potential zones.

Random Forest (RF)

Random Forest is an ensemble learning algorithm based on bagging and decision trees aggregation (Breiman, 2001). The ultimate prediction is learned as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \quad (5)$$

where $h_t(x)$is the prediction from the t^{th}tree and $T$is the total number of trees. RF also provides variable importance measures based on Gini impurity reduction. Random Forest also provides variable importance metrics based on the impurity reduction, that can help in making the factors which control the groundwater dynamics interpretable.

Support Vector Machine (SVM)

Support Vector Machines (SVMs) have been designed to find an optimal separating hyperplane in a high-dimensional feature space (Cortes & Vapnik, 1995):

$$f(x) = \mathbf{w} \cdot \phi(x) + b \qquad (6)$$

where $\phi(x)$is the kernel function. The radial basis function (RBF) kernel was employed:

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \qquad (7)$$

Artificial Neural Network (ANN)

Artificial neural network modelling was performed with multilayer perceptron modelling architecture trained with the use of backpropagation (Haykin, 2009):

$$y = f(\sum_{i=1}^{n} w_i\, x_i + b) \qquad (8)$$

where x_i are inputs, $w_i$are weights, b is bias, and $f$is an activation function (ReLU or sigmoid).

Model Training, Validation, and Evaluation

All the machine learning models were trained with python-based implementations. K fold cross validation was used for hyperparameter tuning and optimize performance. Model accuracy was evaluated using ROC-AUC, accuracy, sensitivity and specificity (Fawcett, 2006). Spatial validation was performed by means of independent groundwater fluctuation data sets. Conducted a sensitivity analysis using manipulative changes on the factors weight systematically and machine learning hyperparameters (Dietterich, 2000). The resultant best suitable model groundwater potential map has been divided into three distinct zones and has been advanced as a decision support tool for sustainable management of the groundwater resources, borehole site selection and basin level planning for the Mahanadi River Basin.

RESULTS AND DISCUSSION

Groundwater Potential Mapping Using AHP

The groundwater potential map by using the Analytic Hierarchy Process (AHP) was applied to integrate twelve groundwater conditioning factors by using weighted overlay analysis. The concomitant groundwater potential index (GWPI) was divided into three areas, namely, poor, moderate, and good groundwater potential zones (Fig. 4). Spatially, delineations of poor to good GPZ occur with a dominance in low to moderate slope terrains, valley-fills, moderate drainage density, high lineament density and favourable geomorphic units, hence implying its enhanced recharge and storage capacities. Conversely, areas of poor potential are associated with structurally less fractured hard rock terrain, steeper topography, and increased drainage density and, therefore, favour runoff over infiltration. Although the AHP model provides a logical and interpretable groundwater potential map, it is limited in its predictive capability by the subjectivity involved in the weighting of the factors as well as the limited ability to represent non-linear interactions among the conditioning variables. Consequently, there is a very compelling need to apply data - driven machine learning methodologies that could potentially help improve predictive accuracy (Kumar et al., 2021; Şahin et al., 2020).

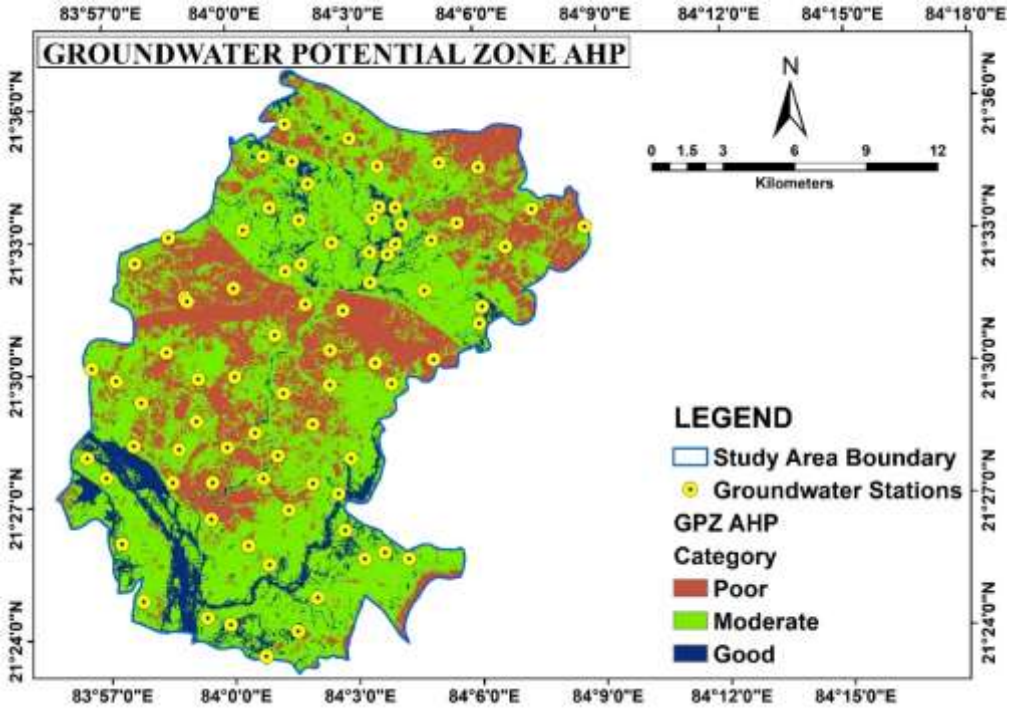

Fig. 4. Groundwater potential map by AHP

Groundwater Potential Mapping Using Frequency Ratio (FR)

The Frequency Ratio (FR) model is used to quantify the groundwater occurrence statistical association to individual conditioning factors. FR analysis showed that lineament density, geomorphology, slope, and drainage density have higher FR values, suggesting high spatial relationship with the occurrence of groundwater. The FR-based groundwater potential map (Fig. 5) shows spatial patterns generally quite in agreement with the AHP output, especially in the low-lying and structurally controlled areas. Nevertheless, the model assumes that conditioning factors are independent, which limits the model's ability to capture complex interactions. Despite this limitation, FR has more favourable performance than the purely knowledge-driven models and can serve as an effective baseline statistical model for groundwater potential assessment. The robust statistical links which the FR model identified between groundwater occurrence and lineament density, geomorphology, and slope. Although FR achieved the highest ROC-AUC value (0.87), its relatively less overall accuracy (0.71) and kappa (0.55) indicate class-wise spatial prediction procedure as suboptimal. These results indicated that using FR both to screening and assessing factor relevance to groundwater potential is satisfactory in consideration of the results; however, the accuracy in prediction value using 'FR' alone is not sufficient and cannot be used alone as a predictive model to zone groundwater potential.

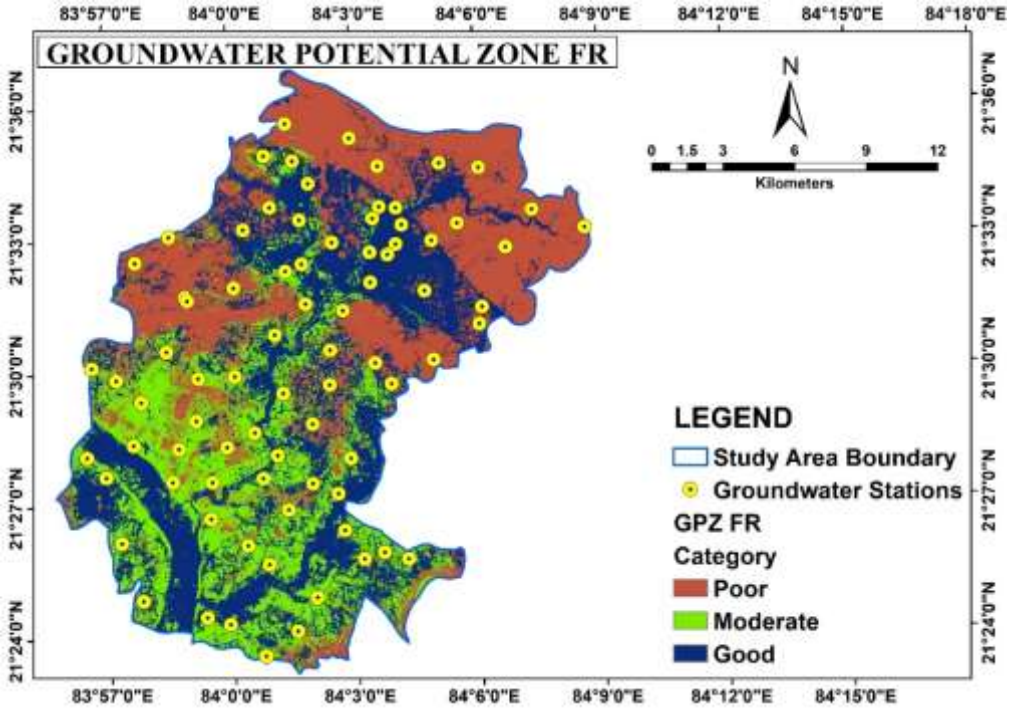

Fig. 5. Groundwater potential map by FR

Groundwater Potential Mapping using Logistic Regression (LR) Model

The Logistic Regression (LR) model was used to assess the linear relationship between the occurrence of groundwater and the selected conditioning factors. The results of the LR analysis suggest that rainfall, lineament density, slope and TWI have a significant influence on the occurrence of groundwater which is indicated by their regression coefficients. The LR- based groundwater potential map (Fig. 6) shows moderate predicted performance. However, the linear nature, assumed by the model, in the logit space limits the model's ability to simulate the complex hydrogeological processes especially in uneven hard rock terrain like that found in the Sambalpur region. Consequently, LR provides lower predictive accuracy as compared to non-nonlinear models of machine learning, but is still useful for interpretability and understanding the directions of influence of factors. Logistic Regression gave a good accuracy value of 0.78 but a smaller value of the Area Under the Curve, 0.68, which indicates the poor ability of this method to model non-co-linear hydrogeological interactions. Nevertheless, a positive effect of rainfall, TWI, and lineament density on the likelihood of the occurrence of groundwater can be assumed from LR due to its valuable interpretability.

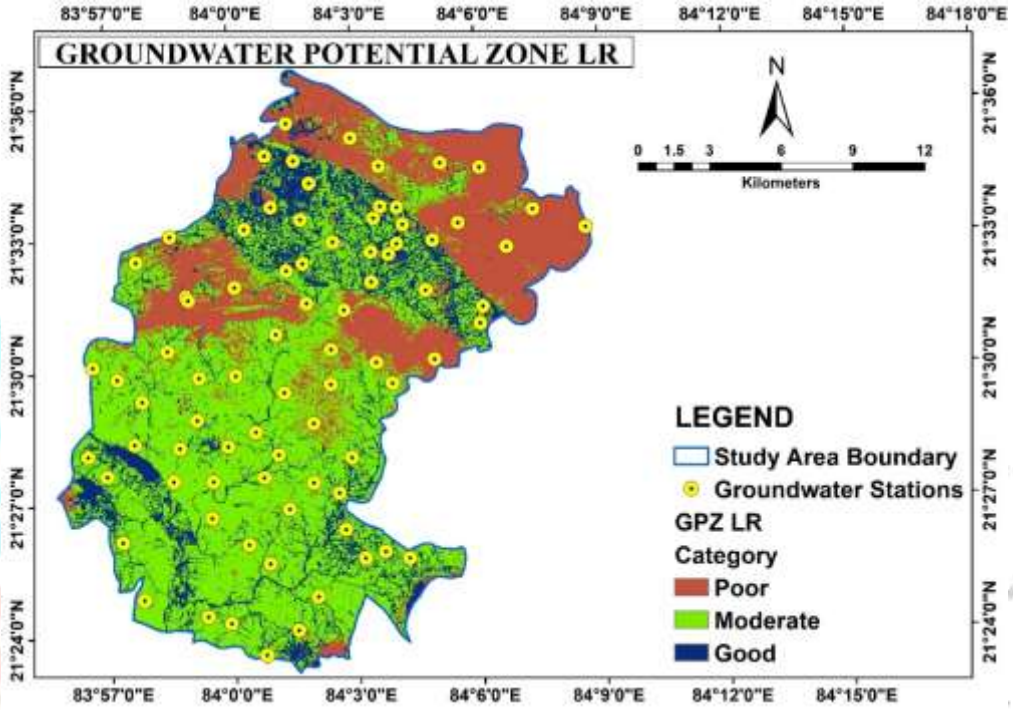

Fig. 6. Groundwater potential map by LR

Groundwater Potential Mapping using Random Forest (RF) Model

The Random Forest (RF) model showed the best predicted results amongst the models in this review as shown by the highest ROC-AUC value (Table 3). The RF-based groundwater potential map (Fig. 7) had very well-defined areas of good potential along structurally controlled and geomorphologically favorable areas. Variable importance analysis shows that for most of them lineament density, slope, geomorphology, drainage density, and rainfall are the most influential predictors. The RF model is able to capture non-linear interdependencies and hierarchy of conditioning factors well that is what makes its performance better. These results support previous studies, which identified RF as one of the best models to measure groundwater potential in complex terrains. The accuracy of the RF model was high (0.91) and so was the agreement in kappa-statistics (0.86). Its capacity to model non - linear interactions with capacity to rank the variable importance, makes it one of the most reliable model for groundwater potential mapping in Sambalpur region. RF is always used to delineate good groundwater potential phenomena in structurally controlled and/or low relief territories, in agreement with the observed groundwater conditions.

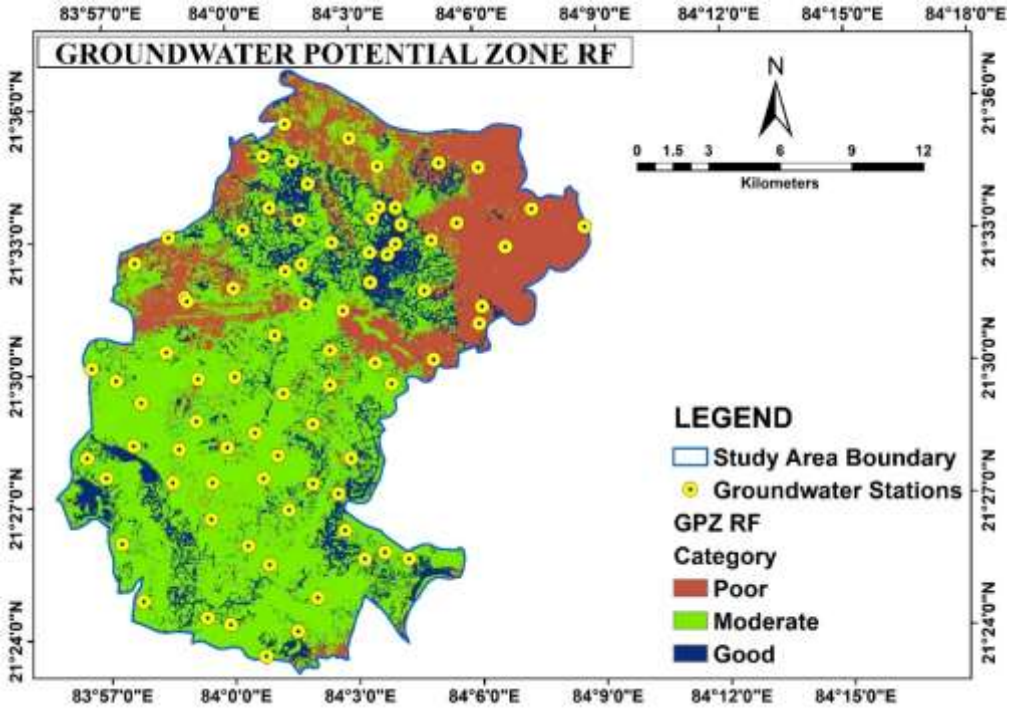

Fig. 7. Groundwater potential map by RF

Groundwater Potential Mapping using Support Vector Machine (SVM) Model

The SVM model using the radial basis function kernel was found to obtain reliable groundwater potential prediction results and ROC-AUC values were similar to RF but slightly lower. The SVM output map (Fig. 8) reveals well defined groundwater potential zones, especially in areas with medium terrain complexity. However, the performance of SVM is sensitive to the choice of kernel parameters and data scaling. While SVM is an effective classifier that can separate the groundwater presence and absence classes well in high dimensional feature space, it is not very interpretative when compared to tree-based classifiers. Nevertheless, SVM has proved to be a good classifier in groundwater potential mapping activities provided with adequate training data and parameter optimisation. SVM acquired reasonable accuracy (0.84) and AUC (0.78) and performed adequately in the moderately heterogeneous terrain. Nevertheless, its parameter kernel sensitivities and low interpretability limit wider application.

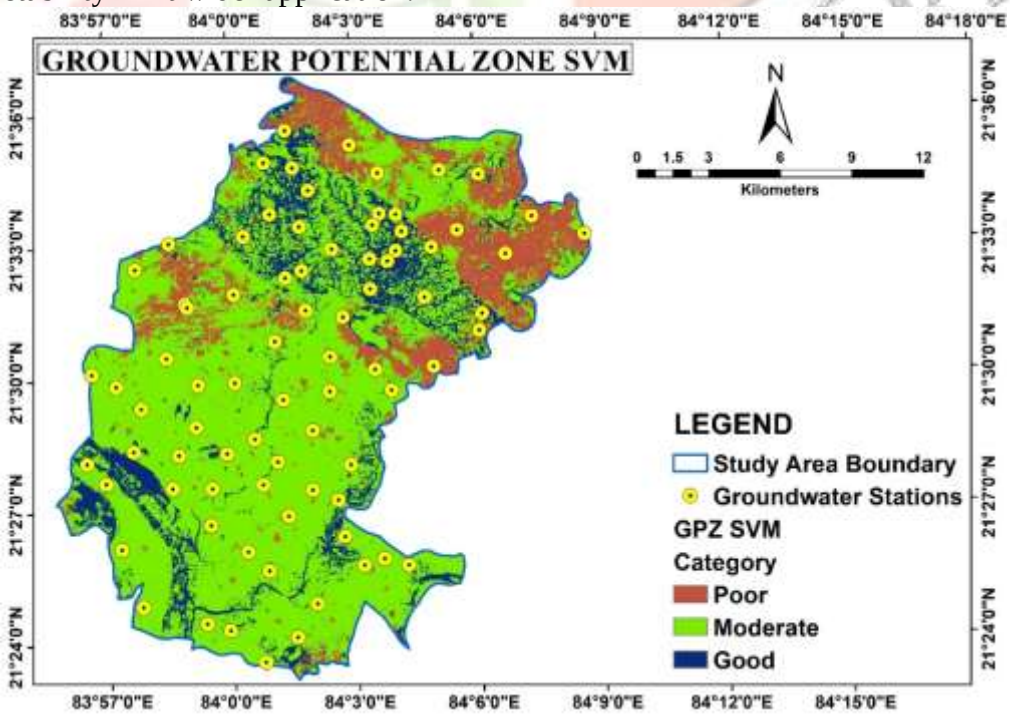

Fig. 8. Groundwater potential map by SVM

Groundwater Potential Mapping using k-Nearest Neighbors (KNN) Model

The KNN model showed a moderate predictive model. Its results show that the prediction of groundwater potential is affected by the proximity of similar hydrogeological situations in space (Fig. 9). However, KNN performance is affected in the areas where spatial heterogeneity is high, as the model is sensitive to noise and k-value choice. The lack of explicit learning process makes them lack generalization capability as compared to ensemble and margin-based classifiers. Thus, KNN is less useful for basin scale groundwater potential mapping but can be used for supplementary information. The KNN model degraded relatively less accuracy (0.76) caused by its sensitivity to the effect of noise and spatial heterogeneity. It is because it is based on instances and therefore there is a limitation for generalization to perform the basin-scale study of groundwater.

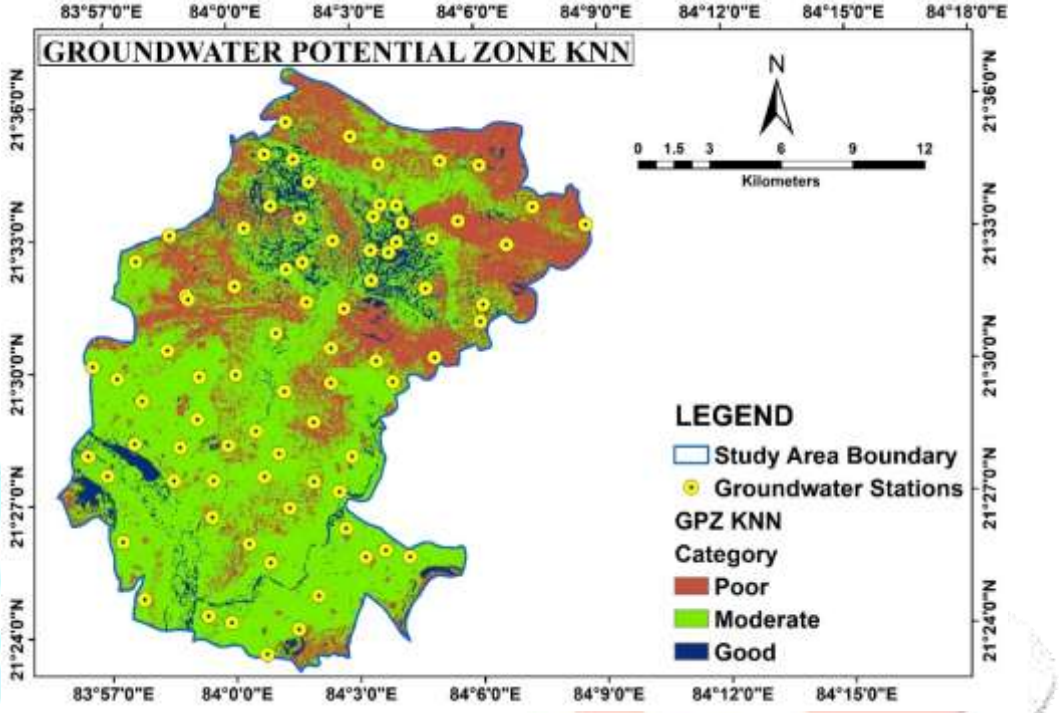

Fig. 9. Groundwater potential map by KNN

Groundwater Potential Mapping using Artificial Neural Network (ANN) Model

The ANN model was able to capture complex non-linear relations between the occurrence of groundwater and conditioning factors. The ANN based groundwater potential map (Fig. 10) shows good potential zones that are in accordance with RF and SVM output. Despite great prediction capability, ANN is a black box model, and has poor interpretability. In addition, the ANN performance is also sensitive to network architecture, learning, and training epochs. ANN is thus useful for prediction purposes, but less useful for interpretation of factors importance and decision transparency. ANN achieved the highest level of accuracy (0.94) and kappa (0.90), which means excellent agreement with ground truth. The model was able to capture complex non - linear interactions among conditioning factors like TWI, rainfall, slope and lineament density. Despite the fact that ANN is a black box, it proved to be the best performing predictive model in this study.

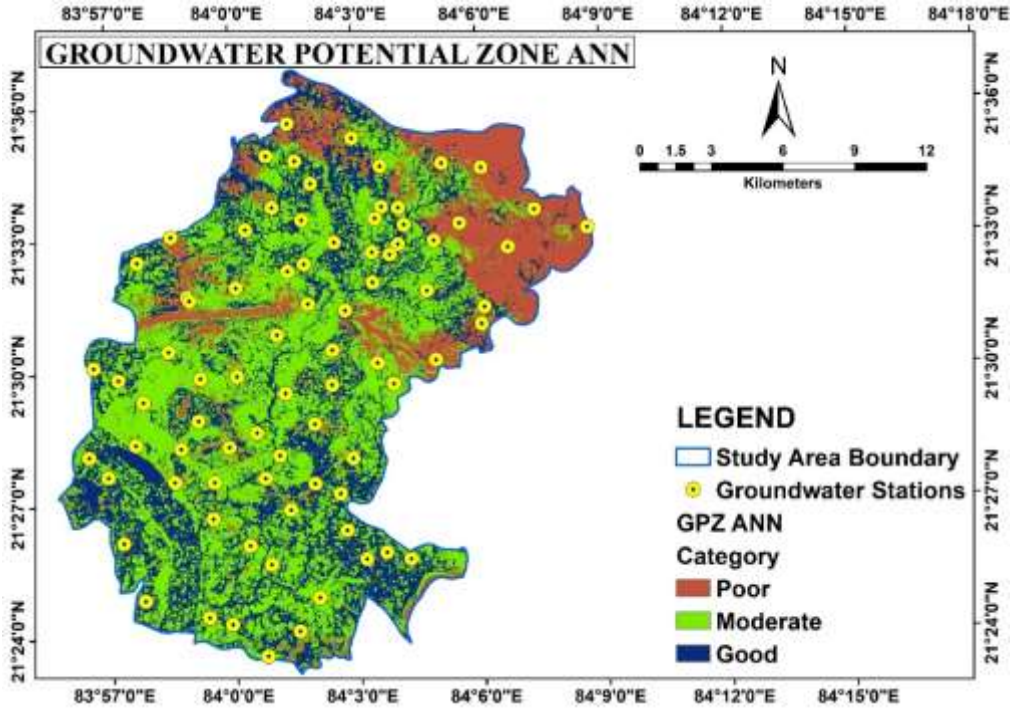

Fig. 10. Groundwater potential map by ANN

Groundwater Potential Mapping using XGBoost Model

The XGBoost model had a performance similar to RF and slightly better in some cases. The boosting framework is adept at minimizing the prediction error and also manages complex interactions between variables. The control of over-fitting thanks to regularization helps XGBoost to improve its ability to generalize. However, computational complexity and parameter sensitivity need to be carefully tuned. The good performance of the model validates the use of gradient-boosting methods for groundwater potential mapping (Fig. 11). XGBoost proved to be on the murmur of 0.89 (accuracy) and 0.79 (AUC). The use of the gradient boosting framework was effective to minimize the misclassification errors and achieving high spatial generalization. However, its computational complexity is more than the RF model.

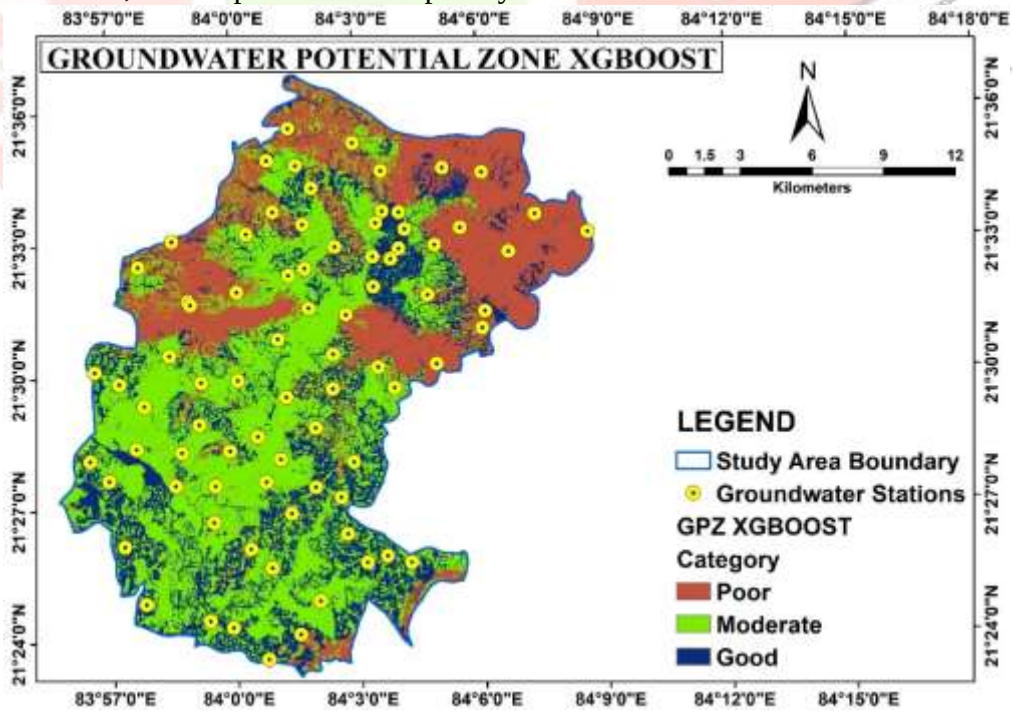

Fig. 11. Groundwater potential map by XGBoost.

Model-wise Performance Summary (GT-based Validation)

Table 3. Comparative performance of groundwater potential models

Model	Accuracy	ROC–AUC	Kappa	Performance Rank
ANN (MLP)	**0.9375**	0.78	**0.90**	**1**
Random Forest (RF)	0.9125	0.82	0.86	2
XGBoost	0.8875	0.79	0.82	3
SVM	0.8375	0.78	0.73	4
Logistic Regression	0.7750	0.68	0.64	5
KNN	0.7625	0.77	0.61	6
Frequency Ratio (FR)	0.7125	**0.87**	0.55	7

ANN shows highest accuracy and kappa, indicating best agreement with ground truth. FR shows high AUC but low accuracy, highlighting that ROC–AUC alone is insufficient. RF and XGBoost are most stable and balanced models. Figure 12 shows the receiver operating characteristic (ROC) curves which represent the discrimination ability of each model for groundwater potential class differentiation. The Frequency Ratio (FR) model achieves the best area under the ROC curve (ROC-AUC=0.87), and thus has a good ranking ability. Nevertheless, its relatively small classification accuracy of 0.71 suggests limited spatial reliability in multi class groundwater potential mapping. On the other hand, the Random Forest (AUC = 0.82) and XGBoost (AUC = 0.79) models show a better and less biased relationship between discrimination and classification performance. The artificial neural network (ANN) is ranked as manifesting a moderate ROC-AUC of 0.78, and also has the best ground truth accuracy of 0.94 underlining its good capacity to represent the complex non-linear interaction. This gap confirms the need to contextualise ROC-AUC with ground truth validation methods and more so in the integration of spatial hydrological studies.

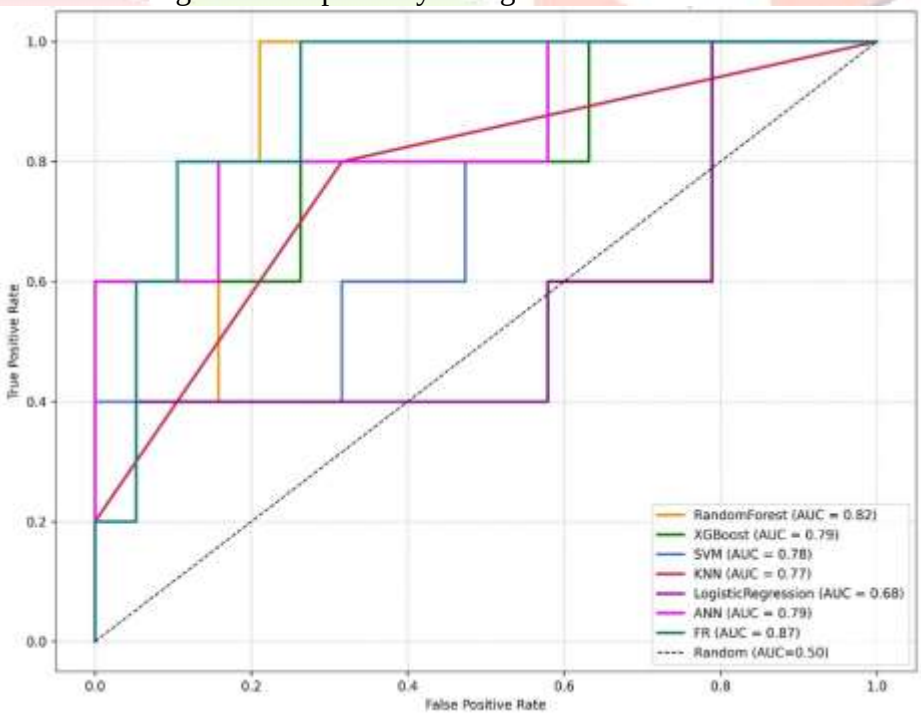

Fig. 12. ROC-AUC Curve of Machine Learning Models

Comparative Model Evaluation

A comparative assessment of all models with respect to ROC-AUC, accuracy and spatial consistency (Table, 3; Figure, 12) shows the following ranking of the models with respect to performance: RF ≈ XGBoost > SVM > ANN > LR > FR > KNN. Tree-based ensemble models continuously outperform statistical and instance-based models because the former are able to model non-linear relationships and interaction between the conditioning factors.

Ensemble Groundwater Potential Map

An ensemble groundwater potential map (Fig.13) was produced by combining the best models. The ensemble output minimizes model specific uncertainty and improves the spatial reliability. Figure 14 shows good accuracy of ensemble machine learning model. The ensemble map has good correspondence with the known groundwater rich areas and CGWB well yield data making it good enough for groundwater resource planning, well site selection and sustainable management. Based on accuracy, kappa and spatial agreement the model performance rankings are as follows ANN > RF > XGBoost > SVM > LR > KNN > FR. An ensemble method which uses ANN and RF as well as the newer technique XGBoost is therefore recommended for the operation of groundwater potential mapping.

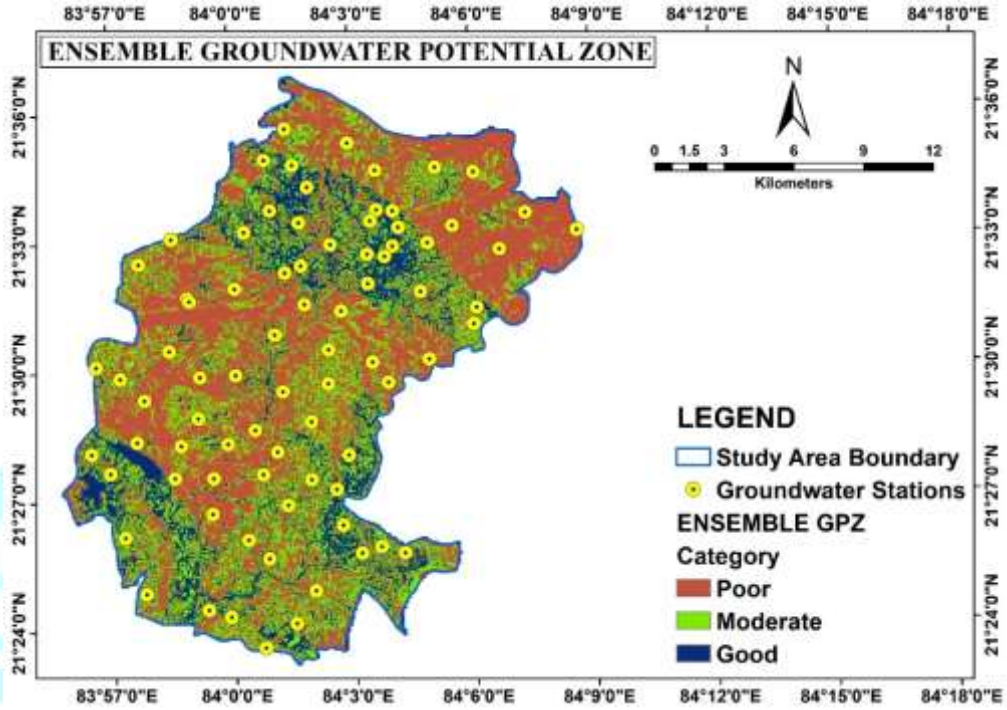

Fig. 13. Ensemble Groundwater Potential Map

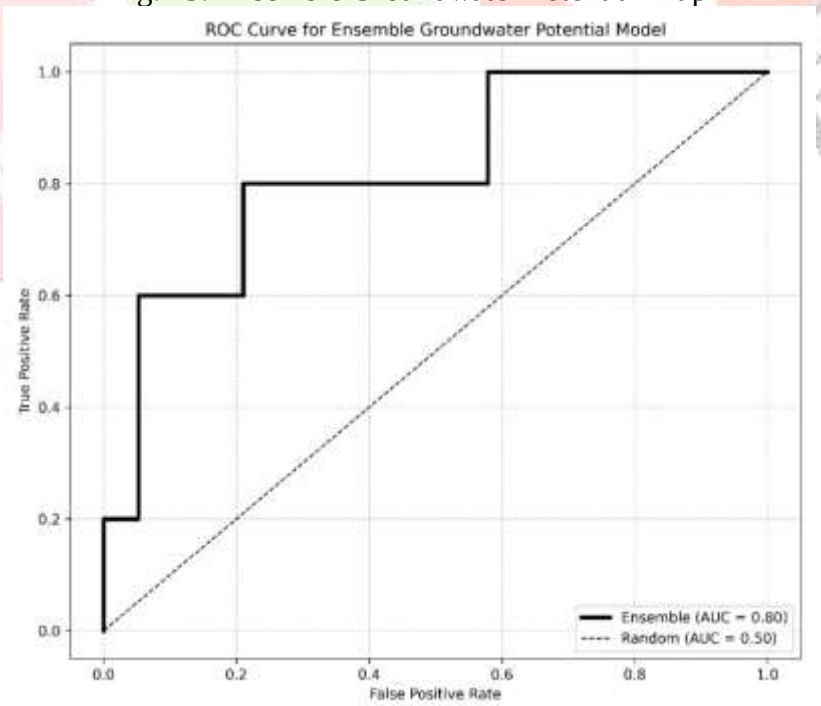

Fig. 14. ROC-AUC Curve of Ensemble model

Implications for Groundwater Management

The results show that groundwater potential assessment in hard rock monsoon dominated basins can be effectively carried based on tools that use GIS and machine learning models such as RF and XGBoost. The identification of the areas of high groundwater potential can help policymakers and planners prioritize groundwater development and at the same time avoid overexploitation.

1. Well-Site Selection
The high groundwater potential zones identified by ANN and RF models can be used to guide the well-siting process scientifically and improve the borewell failure rate.
2. **Sustainable Groundwater Planning**
Moderate potential zones should be given priority for regulated abstraction and artificial recharge; while, in low potential zones, demand management is required.
3. **Decision Support for Local Authorities**
The final groundwater potential map provides opportunities for actionable insights for district-level water resource planning in line with CGWB monitoring strategies.
4. **Climate Resilience**
Incorporating rainfall and terrain indices for better understanding the dynamics of recharge under monsoon variability for climate resilience in groundwater management.
This study shows the significant potential of GIS (integrated with Machine Learning model) especially ANN and Random Forest model, compared to traditional statistical and knowledge-based approach, to map groundwater potential which provides reliable decision support tools for sustainable groundwater management in Mahanadi river basin.

CONCLUSION

The current study affirms the viability of the use of the integrated GIS-AHP-Machine Learning approach in delineated potential groundwater zones in selected parts of the Mahanadi River Basin of Sambalpur district of Odisha. By combining twelve hydrogeologically relevant conditioning variables, along with the evaluation of a set of predictive models to provide an integrated characterization of groundwater potential, it provides a thorough evaluation of groundwater potential in a complex hard-rock and monsoon-dominated hydrologic setting. A comparative evaluation shows that the Artificial Neural Network (ANN) and Random Forest (RF) algorithms have better predictive performances, with the highest ground truth validation result in accuracy and kappa statistics. In relation to this, while the Frequency Ratio model produces a relatively high ROC-AUC model, the low classification accuracy highlights the limitations of having accessibility only to threshold-independent indicators for spatial groundwater evaluation. Tree-based ensemble approaches and neural networks are particularly good at coding nonlinear interactions among groundwater controlling variables and, therefore, make more robust spatial predictions. The resultant groundwater potential maps show that the areas of good potential are mainly located in the setting characterized by low slope, lineament networks, dense lineament networks, favorable geomorphic units, moderate drainage density, and high topographic wetness which corroborates the significant role of terrain and structural controls on the distribution of groundwater within the study domain. From the resource management perspective, the results can be used to provide practical information in scientific well-site selection, groundwater development planning, and identification of recharge zones on a district as well as basin scale. The methodology and findings provide decision-makers, water-management pattern and local accounts with evidence to set up sustainable and spiritually instructable groundwater management strategies. Moreover, the proposed framework shows transferability and could be successfully adapted to other data-poor river basins that have similar hydrogeological frameworks.

REFERENCES

1. Adiat, K. A. N., Nawawi, M. N. M., & Abdullah, K. (2012). Assessing the accuracy of GIS-based elementary multicriteria decision analysis as a spatial prediction tool: A case of predicting potential zones of sustainable groundwater resources. *Journal of Hydrology, 440–441*, 75–89. https://doi.org/10.1016/j.jhydrol.2012.03.028
2. Bonham-Carter, G. F. (1994). *Geographic information systems for geoscientists: Modelling with GIS*. Pergamon Press.
3. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Central Ground Water Board. (2023). *Ground Water Year Book – India 2022–2023*. Ministry of Jal Shakti, Department of Water Resources, River Development & Ganga Rejuvenation, Government of India. https://cgwb.gov.in/cgwbpnm/public/uploads/documents/17135107192062732741file.pdf
5. Central Ground Water Board. (2024). *Dynamic Ground Water Resources of Odisha, 2024*. Ministry of Jal Shakti, Department of Water Resources, River Development & Ganga Rejuvenation,

Government of India. https://cgwb.gov.in/cgwbpnm/public/uploads/documents/17435864691666329414file.pdf
6. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
7. Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.
8. Esri. (2020). *ArcGIS Desktop: Release 10.8*. Environmental Systems Research Institute.
9. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
10. Haykin, S. (2009). *Neural networks and learning machines* (3rd ed.). Pearson.
11. Jari, A., Bachaoui, E. M., Hajaj, S., Khaddari, A., Khandouch, Y., El Harti, A., ... & Namous, M. (2023). Investigating machine learning and ensemble learning models in groundwater potential mapping in arid region: case study from Tan-Tan water-scarce region, Morocco. *Frontiers in Water*, *5*, 1305998.
12. Jha, M. K., Chowdary, V. M., & Chowdhury, A. (2010). Groundwater assessment in Salboni Block, West Bengal (India) using remote sensing, geographical information system and multi-criteria decision analysis techniques. *Hydrogeology Journal, 18*(7), 1713–1728. https://doi.org/10.1007/s10040-010-0631-z
13. Kumar, M. D., & Bassi, N. (2021). The climate challenge in managing water: Evidence based on projections in the Mahanadi River Basin, India. *Frontiers in Water*, *3*, 662560.
14. Kumar, R., et al. (2021). A comparative study of machine learning and Fuzzy-AHP technique to groundwater potential mapping in the data-scarce region. *Computers & Geosciences*, 155, 104855. **https://doi.org/10.1016/j.cageo.2021.104855**
15. 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. https://doi.org/10.1007/s11269-010-9749-y
16. Malakar, P., Mukherjee, A., Bhanja, S. N., Saha, D., Ray, R. K., Sarkar, S., & Zahid, A. (2020). Importance of spatial and depth-dependent drivers in groundwater level modeling through machine learning. *Hydrology and Earth System Sciences Discussions*, *2020*, 1-22.
17. Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management, 31*(9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3
18. Pandey, H. K., Singh, V. K., & Singh, S. K. (2022). Multi-criteria decision making and Dempster-Shafer model–based delineation of groundwater prospect zones from a semi-arid environment. *Environmental Science and Pollution Research*, *29*(31), 47740-47758.
19. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
20. Prasad, P., Loveson, V. J., Kotha, M., & Prasad, M. (2020). Application of machine learning techniques in groundwater potential mapping along the west coast of India. *GIScience & Remote Sensing*, *57*(6), 735–752. https://doi.org/10.1080/15481603.2020.1794104
21. Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2019). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping. *Catena*, *179*, 204–214. https://doi.org/10.1016/j.catena.2019.03.014
22. Raj, D. K., & Gopikrishnan, T. (2024). Optimizing ETo Prediction in the Mahanadi Basin: A Comprehensive Evaluation of Machine Learning Models with Emphasis on ANFIS Performance.
23. Roy, D. K., Hashem, A. A., Reba, M. L., Leslie, D. L., & Nowlin, J. (2024). A maximal overlap discrete wavelet packet transform coupled with an LSTM deep learning model for improving multilevel groundwater level forecasts. *Discover Water*, *4*(1), 16.
24. Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill.
25. Şahin, M., et al. (2020). A comprehensive analysis of weighting and multicriteria methods in the context of sustainable energy. *Water Resources Management*, 34(14), 4395–4412. **https://doi.org/10.1007/s11269-020-02677-4**
26. Scanlon, B. R., Fakhreddine, S., Rateb, A., de Graaf, I., Famiglietti, J., Gleeson, T., ... & Zheng, C. (2023). Global water resources and the role of groundwater in a resilient water future. *Nature Reviews Earth & Environment*, *4*(2), 87-101.
27. Singandhupe, R. B., & Sethi, R. R. (2016). Groundwater resources, its mining and crop planning in Orissa, India. *Int. J. Sci. Res. Publ*, *6*(11), 26-32.

28. Sishodia, R. P., Shukla, S., Wani, S. P., Graham, W. D., & Jones, J. W. (2018). Future irrigation expansion outweigh groundwater recharge gains from climate change in semi-arid India. *Science of the Total Environment*, *635*, 725-740.
29. Vayadande, K., Sadake, S., Sangwai, S., Patil, M., Kadam, S., & Daga, S. (2024). Landslide Susceptibility Prediction System.