Predictive Analytics for Identifying Surface Water Sources for Domestic Water Supply in Phuentsholing, Bhutan

Chimi Wangmo[1], Phurpa Wangmo[2*], Nidup Rinchen[3], Ngawang Choezer[4], Sangey Pasang[5]

[2]Electronic and Communication Engineering Department, College of Science and Technology
[3,4]Engineering Geology Programme, College of Science and Technology, , Royal University of Bhutan
[1,5]Civil Engineering Department, College of Science and Technology, Royal University of Bhutan
Email: chimiwangmo.cst@rub.edu.bt[1], phurpawangmo50@gmail.com[2*], nidup610@gmail.com[3], ngawangc002@gmail.com[4], sangeypasang.cst@rub.edu.bt[5]

Received: 14 April 2025; Revised: 9 June 2025; Acceptance: 10 July 2025; Published: 17 August 2025

Abstract
Surface water is a primary source of drinking water in Bhutan. To ensure a reliable supply that meets both quality and quantity requirements, it is crucial to identify suitable sources. This study presents an integrated approach using Geographic Information Systems (GIS) and machine learning to identify potential surface water sources. Satellite imagery from Landsat-8 and Sentinel-2 was utilized to generate geospatial datasets. Five key variables influencing the spatio-temporal presence of water—rainfall, temperature, soil type, Normalized Difference Vegetation Index (NDVI), and topography were analyzed within a GIS environment. The Random Forest (RF) algorithm, known for its robustness in handling nonlinear and high-dimensional data, was employed to predict potential water sources. Model outputs were validated through field surveys and spectral analysis using the Normalized Difference Water Index (NDWI). The study identified 50 viable water source locations situated above 450 meters in elevation. The model achieved an area under the curve (AUC) score of 0.99, indicating a strong correlation between predicted and actual water sources. These results confirm that integrating machine learning with remote sensing and GIS is an effective approach for surface water resource planning in Bhutan's hilly terrain.
***Keywords**: Surface Water, GIS, NDVI, NDWI, Random Forest*

1. INTRODUCTION

In natural resources management and environmental monitoring, Geographic Information System (GIS), Machine learning and Artificial Intelligence are commonly used tools. These tools are used for Landslide Susceptibility Assessment (Pasang and Kuicek 2020, Shazad et al., 2022), detection of Land Use Land Cover changes (Pasang et la, 2022) and its impacts (Mehra and Swain 2024, Kang et al.,2024), Urban Planning (Anwar et al., 2024) and mapping groundwater sources (Gyeltshen et al., 2022).

In the field of water resources management, the application varies from the use of Geographic Information Systems (GIS) and Remote Sensing techniques to delineate and map surface water bodies (Niu et al, 2022, Pan et al, 2020) and estimate water volume variations (Lin et al, 2020, Pimenta et al, 2024, Quang et al., 2021). It is also used in the assessment of water quality and treatment, energy management, and impact of climate change (Hernandez-Alpizar et al., 2024)

For observation of surface water, water index-and threshold-based approaches are commonly used (Zhou et al., 2017). The normalized difference water index (NDWI), modified normalized difference water index (mNDWI) are widely used to define and improve water detection by using Near-Infrared (NIR) radiation and Middle-Infrared (MIR) respectively (Quang et al., 2021). Other water indices used in identifying the surface water bodies, includes: tasseled cap wetness index (TCW), Sum457, automated water extraction index (AWEI). Advances in satellite imagery and remote sensing enable processing of intricate environmental data.

Further the integration of satellite images (Landsat-8 and Sentinel-2) and machine learning model are found to be useful to generate predictive maps of potential surface water and to process intricate environmental data to solve actual resource management issues (He et al., 2024; Mohan et al., 2025; Zhu et al., 2022; Mohammed et al., 2023).

Surface water sources, particularly springs and streams, are the primary sources of domestic water supply in Bhutan (Tariq et al.,2021). Despite the country's abundant water resources, with an annual per capita availability of 94,500

m^3 (Yangzom and Choden, 2021), Bhutan faces significant challenges in ensuring reliable and equitable water access. These challenges are exacerbated by climate change and human activities, leading to spatio-temporal variations in water quantity and posing difficulties in identifying and locating water sources due to the country's rugged topography. In particular, the water supply system in Phuentsholing, managed by Phuentsholing Thromde, draws water from streams and rivers intakes of the upper reaches of the town as well as ground water sources. However, the water supply system faces several challenges including high seasonal variability, aging distribution networks, and growing demand. The instances of interrupted water supply are common occurrences affecting several households including the residents of the College of Science and Technology.

With changes in the land use pattern, increasing human activities and climate change, drying up of water sources, changes in the water storage and water pollution (Lin et al, 2020) are expected in the catchment area of the Phuentsholing town. This will lead to further pressure on the already stressed water supply system of the town.

Therefore, there is a need to find a sustainable solution to the water shortage particularly during the dry periods, by identifying reliable water sources such as streams, ponds, and ephemeral water bodies which is a critical step towards increasing water security in the region (Roy et al., 2020). However, thus far there has been no assessment carried out to determine the spatio-temporal location of the water sources in the area. Therefore, the aim of the study is to apply predictive analytics and machine learning (ML) to identify surface water sources in the broader zone of Phuentsholing by leveraging geospatial technologies.

2. STUDY AREA

Phuentsholing town, is an important commercial hub and gateway for trade with India and other trading partners. It is located in the southern Bhutan under Chukha District at approximately 26.851°N and 89.388°E (Fig. 1) at an altitude of about 300 meters above sea level.

The town is characterized by subtropical climate with heavy monsoon rain, complex topography, and intensively diversified land use pattern (National Land Commission Secretariat, 2021). The town uses mainly the surface water sources for domestic water supply supplemented by groundwater sources. Two major river system are located within the vicinity of the town; Amochu runs along the periphery of the town while Omchu, a smaller tributary runs through the town.

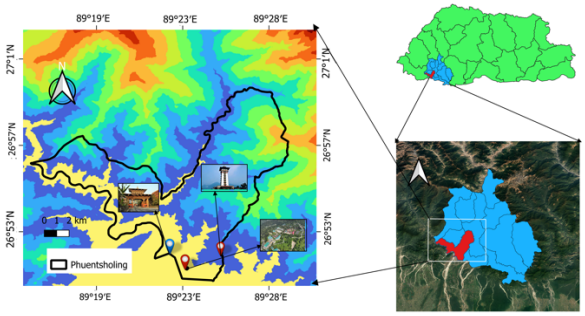

Fig. 1: *Map Showing study area Phuentsholing within the district of Chukha in Bhutan*

Soil type in Phuentsholing dictates water retention, infiltration, and surface water body distribution. The region has diverse types of soil (Fig. 2), each with their own properties that affect hydrology.

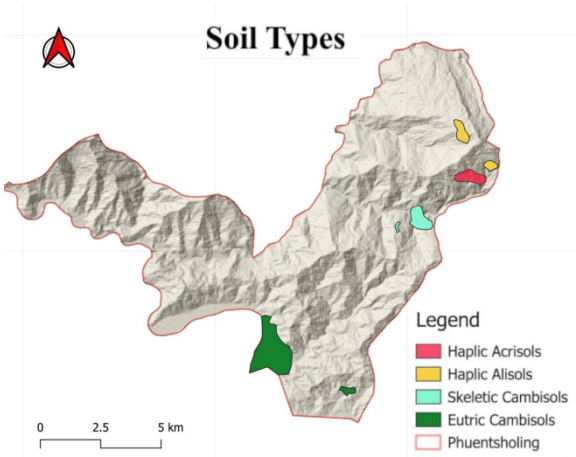

Fig. 2: *Soil Map of study area showing Haptic Acrisols, haptic Alisols, Skeletal Cambisols and Eutric Cambisols*

The interplay between soil properties and topography significantly influences surface water dynamics in Phuentsholing. The region features diverse soil types distributed across varying landscapes, from the flat Indo-Bhutan border plains to undulating hills. Haptic Acrisols and Alisols dominate the humid zones. Both are acidic and clay-rich with low permeability, resulting in limited infiltration and increased surface runoff during heavy rainfall. Haptic Alisols, which contain higher concentrations of exchangeable aluminum, are particularly prone to nutrient depletion and poor groundwater recharge.

In contrast, Skeletic Cambisols, found predominantly on steep slopes, are shallow and coarse-textured. These soils allow rapid infiltration but possess low water storage capacity, contributing to high runoff and limited water retention. Eutric Cambisols, prevalent in valley bottoms and low-lying plains, are fertile and well-balanced in drainage and retention, making them well-suited to support seasonal streams, ponds, and localized water accumulation.

These soil-topographic relationships demarcate distinct hydrological zones. Eutric Cambisols in the lowlands tend to form natural reservoirs, while Skeletic Cambisols on elevated slopes contribute to rapid drainage. Seasonal monsoonal rainfall intensifies these patterns: heavy rain temporarily saturates clayey soils like Acrisols, whereas prolonged dry spells quickly deplete su rface moisture in poorly retaining areas.

Urbanization further alters these natural dynamics. Impervious surfaces, particularly in central Phuentsholing, reduce infiltration and increase dependency on engineered drainage systems. Conversely, vegetated areas enhance soil water retention, especially where Eutric Cambisols are present. These findings underscore the critical role of land use planning in maintaining water sustainability.

Temperature changes also play a critical role, with high rates of evaporation in the hot pre-monsoon period drying up the surface water resources, especially in Skeletic Cambisols regions.

Understanding these complex interactions is crucial in addressing Phuentsholing's water shortage. By incorporating soil data, topography, and climatic data into machine learning models, it becomes possible to predict sustainable surface water sources for the entire district, beyond localized solutions for institutions like CST. This holistic approach can be employed to develop enhanced water management plans, ensuring long-term sustainability against shortages.

3. MATERIALS AND METHOD

The methodology employed to predict the spatial location of surface water sources in the Phuentsholing region based on remote sensing, geospatial analysis, and integration of machine learning is shown in Fig. 3.

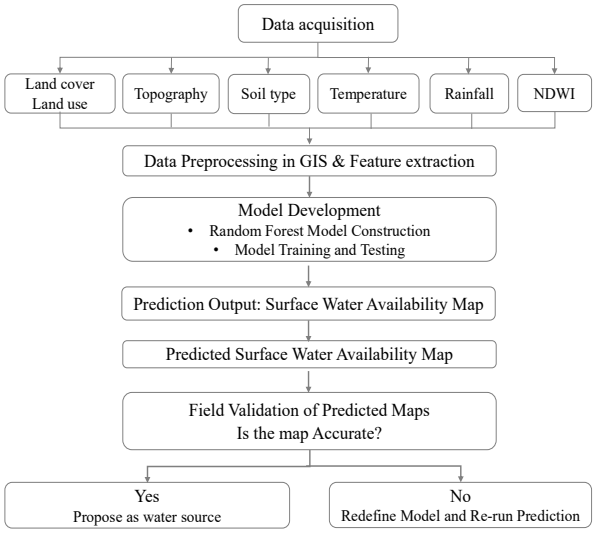

Fig. 3: *Predictive Analysis workflow for surface water presence*

With reference to the Fig. 3, the methodology adopted involves collecting data (soil, temperature, rainfall, land use), preprocessing with GIS tools, and integrating features into a Random Forest model to generate a water prediction map. The map is validated through field visits and statistical techniques with refinements made to generate a spatially consistent and temporally valid surface water prediction map to support water resource planning and management in the region.

3.1. Data Collection and Pre-processing

Six key environmental and geospatial variables were selected for their relevance to surface water distribution: rain, temperature, soil, Normalized Difference Vegetation Index (NDVI), topography, and the Normalized Difference Water Index (NDWI). These variables are readily available from national and global datasets such as the Bhutan Meteorological Department, FAO soil data, and DEM-derived topography. Among them, NDWI is noteworthy as it utilizes spectral characteristics of satellite imagery to precisely detect water bodies by highlighting the contrast between green reflectance and near-infrared (NIR) values (Huang et al., 2018; Zhou et al., 2017; Pan et al., 2020).

The spatial and environmental data were gathered from 2023 to 2025 for a period of three years are shown in figures 5-8. To achieve precision and uniformity, NDWI and NDVI were derived from Sentinel-2 and Landsat-8 satellite images. Both NDWI and NDVI were selected on the basis of their sensitivity towards surface water and vegetation. Data were collected from

the dry season months of each year, i.e., January, February, November, and December. The months were chosen to capture more permanent surface water bodies, minimizing the effect of temporary flooding or seasonal water accumulation that occurs during the monsoon.

Data processing was carried out in GIS platform to prepare, normalize, and spatially register the data sets. Randomly selected points (Fig. 4) were located throughout the Phuentsholing area to maximize spatial diversity and minimize sampling bias. At each station, reflectance values from the satellite data were used to find the NDWI and NDVI values. The NDWI was computed using the formula:

$$\text{NDWI} = \frac{(\text{Green} - \text{NIR})}{(\text{Green} + \text{NIR})}$$

While NDVI was obtained using the following equation:

$$\text{NDVI} = \frac{(\text{NIR} - \text{Red})}{(\text{NIR} - \text{Red})}$$

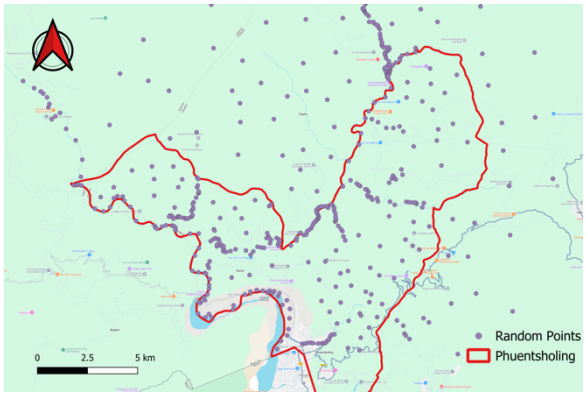

Fig. 4: *Map showing 500 randomly picked points used for in analysis*

All data were pre-processed with standard pre-processing procedures including cloud masking, normalization, projection alignment, and raster-to-tabular conversion for machine learning purposes. Topographic factors were computed using Digital Elevation Models, and climatic data were extracted from national databases. Soil maps and vegetation were also included in the dataset. The result was an integrated geospatial database with a number of variables influencing the occurrence of surface water in the study area.

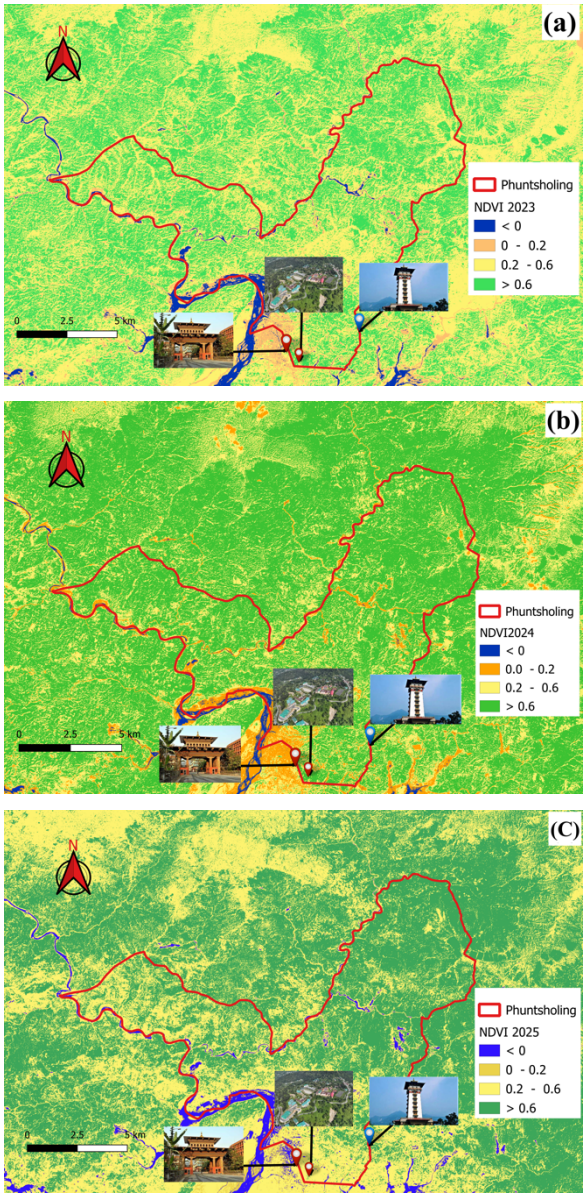

Fig. 5*. Sample NDVI Maps of (a) January 2023, (b) November 2024 and (c) January 2025*

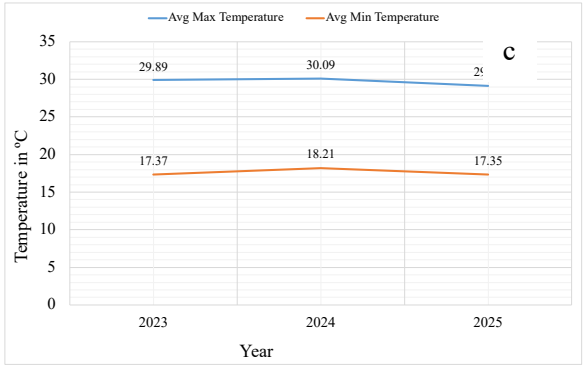

Fig. 6: *Annual temperature in °C in Phuentsholing from 2023-2025*

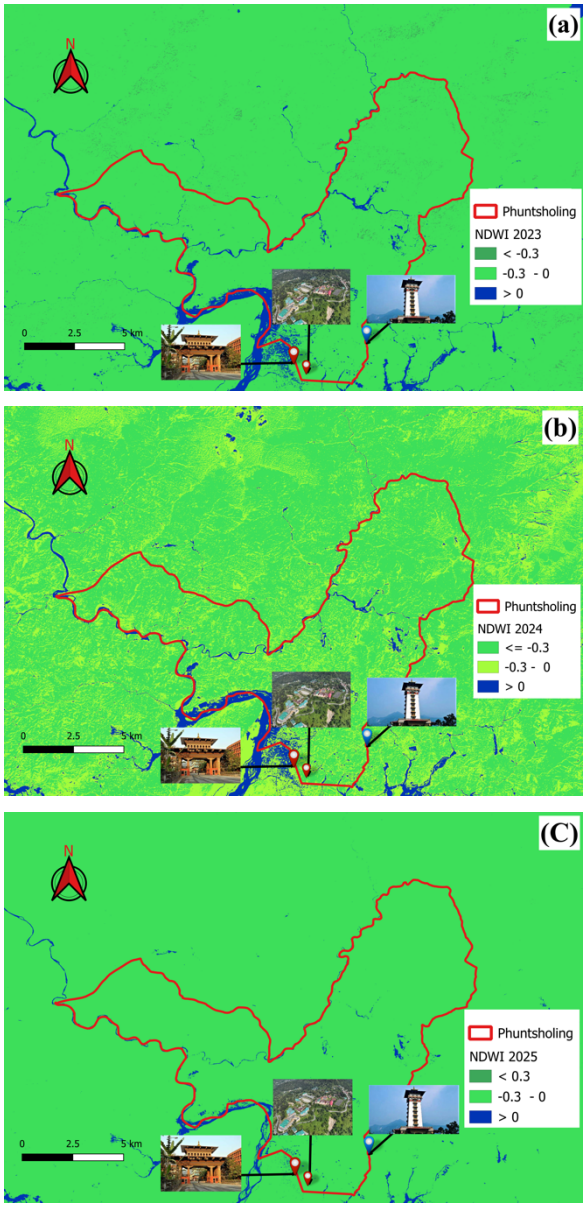

Fig. 7: *Sample NDWI Maps of (a) January 2023, (b) December 2024 and (c) January 2025*

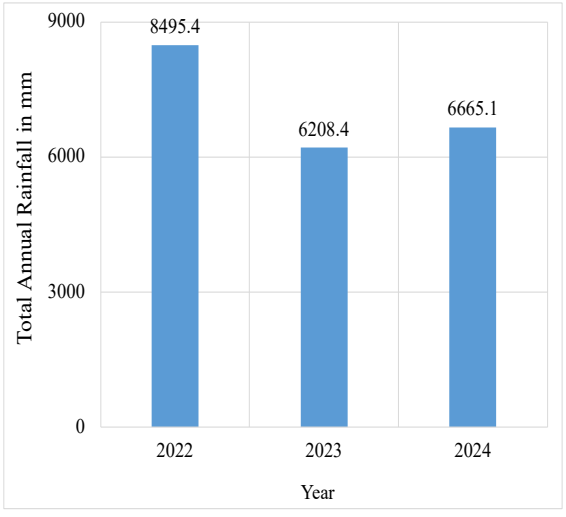

Fig. 8: *Annual rainfall data from 2023-2024*

3.2. Development of Random Forest Model

Spatial data processed in the GIS platform was used for model development using the Random Forest algorithm. Random Forest was adopted as the primary modeling technique due to its effectiveness in handling high-dimensional datasets, resistance to overfitting, and robustness in capturing complex nonlinear relationships. Further, Random Forest is supported by other ensemble machine learning methods including Gradient Boosting and AdaBoost, which also undergo testing in order to be compared on a performance basis (Kathirvelu et al., 2023; Rahaman et al., 2023). Random Forest also provides insights into feature importance, enabling the identification of variables that contributed most significantly to the model's predictions

The model was trained on a merged dataset, where each observation represented a unique geographic location characterized by a set of environmental features and a binary indicator denoting the presence or absence of surface water. Key hyperparameters such as the number of trees, maximum tree depth, and minimum samples per split were specified and tuned to optimize model performance.

To enhance model reliability and mitigate overfitting, the dataset was partitioned into training and validation sets using an 80:20 split. Additionally, five-fold cross-validation was employed on the training set to assess the model's generalizability. In this process, the training data was divided into five equal subsets; in each iteration, the model was trained on four subsets and validated on the remaining one. This procedure was repeated five times, ensuring that each subset served as the validation set once. The results from all folds were averaged to evaluate overall model performance

3.3. Model Validation

Validation of the model's performance was conducted using a combination of statistical evaluation and field validation. Statistical validation was carried out by calculating the performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics were utilized to assess the performance of the model in correctly classifying the presence and absence of surface water across the validation dataset.

Field validation of estimated surface water locations was performed by comparing the locations with ground observations made through site visits in the Phuentsholing area. High surface water probability areas were mapped and assessed to validate the presence of water features. Hand-held GPS devices and mobile GIS applications were used to locate and note water features were used during the field visit. Where areas were not accessible, high-resolution imagery available on platforms like Google Earth was used to visually detect water presence.

4. RESULTS AND DISCUSSION

The final predictive model identified 50 potential surface water bodies as shown in **Error! Reference source not found.**. These sites are located at an elevation above 450 meters, which is the reference elevation of Phuentsholing town. This elevation criterion was purposely applied to facilitate conveyance of water under gravity to the town area.

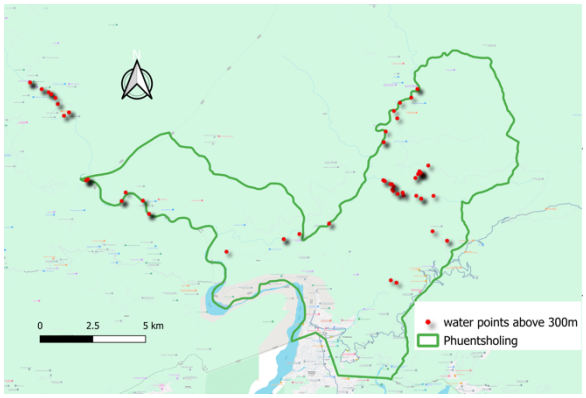

Fig. 9: *Map showing 50 predictive surface water bodies superimposed on the elevation layer*

The input parameters used in prediction model included rainfall, temperature, soil type, NDVI, NDWI, slope, and elevation. NDVI and NDWI indices were derived from Sentinel-2 and Landsat-8 satellite images, while the topographic and climatic variables were processed using Digital Elevation Models and national meteorological data respectively.

4.1 Model Validation (AUC Curve)

The model scored an Area Under the Curve (AUC) of 0.99, signifying a high level of classification accuracy in discriminating between areas with and without surface water presence. Having an AUC value close to 1.0 means that the model is highly capable of ranking positive cases above negative cases. This high AUC value verifies the model as robust and credible in the context of surface water prediction. Therefore, the Random Forest model utilized in this study can be considered valid and effective for identifying prospective surface water sources for the Phuentsholing region.

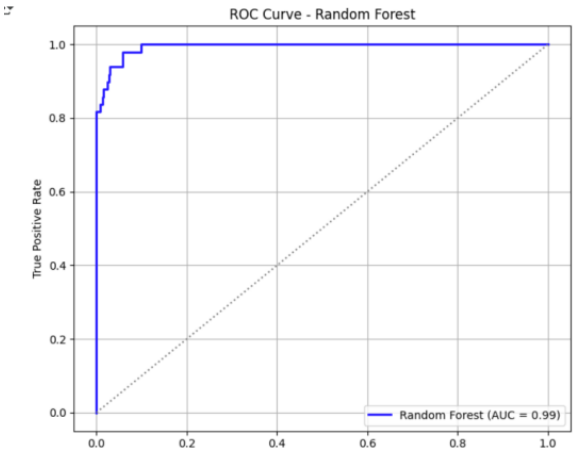

Fig. 10*: The model achieved an Area Under the Curve (AUC) score of 0.99, indicating excellent classification accuracy. The steep rise near the y-axis reflects a high true positive rate with minimal false positives, confirming the model's strong predictive capability*

4.2 Field Validation

To validate the model's predictions, a field visit was conducted to a site identified by the Random Forest model as having a high likelihood of surface water presence. The selected location, situated at coordinates 89.41740552° E and 26.88614462° N and at an elevation above 450 meters (**Error! Reference source not found.**), was accessible during the dry season. On-site verification confirmed the presence of surface water, aligning with the model's prediction and supporting its reliability.

Fig. 11: *Figure showing field validation site. This is where the model selected as having high surface water occurrence probability*

This in-field verification adds validity and consistency of the Random Forest model. While a single point was inspected in the field, the positive result confirms the model's suitability for application on more widespread similar terrain.

***Fig. 12**: Field evidence confirms the presence of surface water at the site identified by the Random Forest model*

5. CONCLUSION

This study investigated and substantiated the application and effectiveness of machine learning and geospatial technologies for delineating potential surface water sources in Phuentsholing region. With data-driven measurements, including NDWI, NDVI, elevation, rainfall, temperature, slope, and soil, the Random Forest model predicted 50 potential surface water locations above the elevation threshold of 300 meters. The predicted results were spatially confirmed by field observation and geospatial overlay, validating the robustness of the model

This approach is distinct from other ongoing research in that it utilizes an ensemble learning model (Random Forest) with a high reputation for accuracy in handling nonlinear environmental data. Utilization of imagery from dry-season months (December, February, January, November) enhanced the model's performance in identifying more permanent surface water bodies free of seasonal bias induced by monsoon rains. Utilization of GIS with machine learning versus traditional field-survey methods or purely hydrological models substantially reduced the demands of physical fieldwork while optimizing spatial coverage. The use of Sentinel-2 and Landsat-8 images added spectral precision to the classification process, especially through NDWI's sensitivity to water reflectance. However, there were a few limitations. Satellite imagery on cloudy days still needed to be passed through cloud-masking processes, and dense canopy or shadowed topography may have obscured surface water features. While the model withstood statistical verification and field validation well, accessibility problems in certain locations limited complete on-site validation. In addition, due to resolution constraints in some of the datasets, smaller water bodies or ephemeral sources may have been missed or misclassified.

For future research, the utilization of imagery of higher resolution or LiDAR-derived elevation models is recommended for more precise outcomes, especially in urbanized or forested regions. Expanding the extent of the model to account for seasonal variation through the inclusion of monsoon datasets can yield information on episodic or temporary water bodies. The inclusion of socio-economic indicators, such as land use patterns or population density, would also be beneficial to rank water supply development areas. This study provides a replicable and scalable model that can be duplicated in other water-scarce regions of Bhutan and beyond.

The findings hold significant implications for Bhutan, which is increasingly water-deficient, especially in lowland urban centers like Phuentsholing. The use of geospatial intelligence in water resource planning provides a cost-effective and scalable alternative to traditional survey methods. It provides policymakers and engineers implementing infrastructure and water supply development with timely, evidence-based recommendations. Having all the forecasted water sources above the elevation of the city also aligns with the realities of logistics of water collection and distribution.

In the future, future studies can leverage the use of temporal analysis that tracks water availability on a seasonal basis. High-resolution spatial data, integration with socio-economic variables, and local stakeholder-based participatory GIS can render such models more viable and useful. This study gives a good platform for predictive water source identification and can inform long-term water security planning in Bhutan.

REFERENCES

Anwar, M. R., & Sakti, L. D. (2024). Integrating artificial intelligence and environmental science

for sustainable urban planning. *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, 5(2), 179-191.

Gyeltshen, S., Kannaujiya, S., Chhetri, I. K., & Chauhan, P. (2022). Delineating groundwater potential zones using an integrated geospatial and geophysical approach in Phuentsholing, Bhutan. *Acta Geophysica*, *71*(1), 341–357. https://doi.org/10.1007/s11600-022-00856-x

He, M., Qian, Q., Liu, X., Zhang, J., & Curry, J. (2024). Recent progress on surface water quality models utilizing machine learning techniques. *Water*, *16*(24), 3616. https://doi.org/10.3390/w16243616

Huang, C., Chen, Y., Zhang, S., & Wu, J. (2018). Detecting, extracting, and monitoring surface water from space using optical sensors: A review. *Reviews of Geophysics, 56*(2), 333–360. https://doi.org/10.1029/2018RG000598

Kathirvelu, K., Yesudhas, A. V. P., & Ramanathan, S. (2023). Spectral unmixing based random forest classifier for detecting surface water changes in multitemporal pansharpened Landsat image. *Expert Systems with Applications,224*,120072.https://doi.org/10.1016/j.eswa.2023.120072

Mohan, S., Kumar, B., & Nejadhashemi, A. P. (2025). *Integration of machine learning and remote sensing for water quality monitoring and prediction: A review*. *Sustainability,17*(3),998.https://doi.org/10.3390/su17030998

Mohammed, M. A. A., Kaya, F., Mohamed, A., Alarifi, S. S., Abdelrady, A., Keshavarzi, A., Szabó, N. P., & Szűcs, P. (2023). Application of GIS-based machine learning algorithms for prediction of irrigational groundwater quality indices. *Frontiers in Earth Science, 11*, 1274142. https://doi.org/10.3389/feart.2023.1274142

Mehra, N. and Swain, J.B. (2024) Assessment of Land Use Land Cover Change and Its Effects Using Artificial Neural Network-Based Cellular Automation. *Journal of Engineering and Applied Science*, 71, (70). https://doi.org/10.1186/s44147-024-00402-0

National Land Commission Secretariat. (2021). Geographical Data of Bhutan. Thimphu, Bhutan: NLCS Publications.*Safeguarding Bhutan's water in the face of climate change | UNDP Climate Change Adaptation*. (2023, May 9).

Pan, F., Xi, X., & Wang, C. (2020). A comparative study of water indices and image classification algorithmsfor mapping inland surface water bodies using Landsat imagery. *RemoteSensing,12*(10),1611. https://doi.org/10.3390/rs12101611

Pasang, S., & Kubíček, P. (2020). Landslide susceptibility mapping using statistical methods along the Asian Highway, Bhutan. Geosciences, 10(11), 430. https://doi.org/10.3390/geosciences10110430

Pasang, S., Norbu, R., Timsina, S., Wangchuk, T.,& Kubicek, P. (2022). Normalized difference vegetation index analysis of forest cover change detecting in Paro Dzongkhag, Bhutan. *Computers in Earth and Environmental Sciences.* (pp. 417-425). Elsevier. https://doi.org/10.1016/B978-0-323-89861-4.00045-2

Rahaman, M. H., Roshani, Masroor, M., & Sajjad, H. (2023). Integrating remote sensing derived indices and machine learning algorithms for precise extraction of small surface water bodies in the lower Thoubal river watershed, India. *Journal of Cleaner Production, 422*, 138563. https://doi.org/10.1016/j.jclepro.2023.138563

Roy, S., Robeson, S. M., Ortiz, A. C., & Edmonds, D. A. (2020). Spatial and temporal patterns of land loss in the Lower Mississippi River Delta from 1983 to 2016. *Remote Sensing of Environment, 250*, 112046. https://doi.org/10.1016/j.rse.2020.112046

Yangzom, K.., & Choden, P. (2021). Climate Change and Water Resources in Bhutan. *Journal of the Bhutan Ecological Society.*

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health, 1*, 107–116. https://doi.org/10.1016/j.eehl.2022.06.001

Zhou, Y., Dong, J., Xiao, X., Xiao, T., Yang, Z., Zhao, G., Zou, Z., & Qin, Y. (2017). Open surface water mapping algorithms: A comparison of water-related spectral indices and sensors. *Water, 9*(4), 256. https://doi.org/10.3390/w9040256