Spatial Modeling of Groundwater Level in Bangladesh Using Physio-Climatic Variables: Machine Learning and Statistical Approaches

A S M Maksud Kamal (✉ maksudkamal@du.ac.bd)  
University of Dhaka

Abul Kashem Faruki Fahim  
University of Dhaka

Shamsuddin Shahid  
Universiti Teknologi Malaysia

Article

Keywords: GLDAS, GRACE, BWDB, Groundwater, Trends, ANN, Bangladesh

Posted Date: November 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2261689/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Spatial Modeling of Groundwater Level in Bangladesh Using Physio-Climatic Variables: Machine Learning and Statistical Approaches

A S M Maksud Kamal\textsuperscript{1*}, Abul Kashem Faruki Fahim\textsuperscript{1} and Shamsuddin Shahid\textsuperscript{2}

\textsuperscript{1}Department of Disaster Science and Climate Resilience, University of Dhaka, Dhaka-1000, Bangladesh (email: maksudkamal@du.ac.bd; fahim-4-2015117896@dsm.du.ac.bd)

\textsuperscript{2}Department of Water & Environmental Engineering, School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia (email: sshahid@utm.my)

*Corresponding author. Email: maksudkamal@du.ac.bd
Spatial Modeling of Groundwater Level in Bangladesh Using Physio-Climatic Variables: Machine Learning and Statistical Approaches

Abstract

Groundwater monitoring is essential for sustainable groundwater resource management in a country like Bangladesh, where this precious resource is gradually declining due to overextraction. Acquiring groundwater level (GWL) over a large area is time-consuming and expensive. This study proposes an alternative approach to groundwater monitoring using freely available daily groundwater storage (GWS) gridded data of the Global Land Data Assimilation System (GLDAS) with other freely available data, including population, rainfall, temperature, irrigation, elevation for modeling GWL data of Bangladesh with a spatial resolution of 0.25° × 0.25°. This was accomplished by employing multiple linear regression (MLR) and artificial neural networks (ANN), using weekly in-situ GWL data at 844 locations distributed over Bangladesh. The results showed the inability of GWS data to estimate the country's groundwater spatial variability and trend. The relative performance of MLR and ANN models revealed a higher capability of ANN in estimating GWL from GWS and other data with an overall correlation coefficient (R) of 0.95 and mean squared error (MSE) of 0.64. The study revealed population and rainfall have the most decisive influence in determining GWL. The model developed using ANN can be used to estimate GWL at locations where observation data are unavailable and thus monitor GWL for sustainable groundwater management.

Keywords: GLDAS; GRACE; BWDB; Groundwater; Trends; ANN; Bangladesh.
1 Introduction

Bangladesh, one of the most populated countries globally, heavily depends on groundwater.\(^1\) In the country, roughly 32 km\(^3\) of groundwater is extracted annually, which is nearly 4% of global groundwater withdrawal.\(^2\) Nearly 90% of abstracted water is used for irrigation. The rapid expansion of irrigated agriculture caused a sharp rise in irrigation demand and, thus, groundwater overexploitation.\(^3\) Groundwater overexploitation caused several hydrological and socioeconomic issues, including water-level declines, water-quality degradation, well drying up, reduction of pumpage in household water supply wells, increased pumping costs, subsidence of land surface, and aquifer compaction.\(^4\) Sustainable groundwater resource management is necessary for the country to ensure future water security. Groundwater abstraction based on sustaining groundwater levels is vital for this purpose.\(^5\) Regular groundwater level observations for the entire country are required to achieve this goal. This necessitates an easily available, feasible, and dependable data source or a model for accurate and reliable groundwater level estimation.

Bangladesh Water Development Board (BWDB) is in charge of regulating and monitoring both the surface and groundwater supplies of Bangladesh. Since 1965, BWDB has maintained a network of more than 1300 groundwater observation wells. However, BWDB data are not freely available. It is expensive for a larger area and a longer period. Besides, the data are not ready to use; data collection, compilation, processing, selecting reliable data from hydrographs, and filling in missing values are all time-consuming.

Global Land Data Assimilation System (GLDAS) gridded groundwater storage (GWS) data may be preferable to a costly groundwater monitoring network. Using advanced data assimilation methods and land surface modelling, the GLDAS considers observational data from the ground and from satellites to produce ideal fields of land surface fluxes and states.\(^6\) It provides global modeled earth surface and atmospheric data at 0.25° × 0.25° and 1° × 1° spatial resolution from 1948 to the present. CLSM GLDAS-2.0 simulates shallow groundwater,\(^7\) which is freely or readily available. Studies in different regions showed its capability as an alternative to in-situ data.\(^8\)–\(^14\) However, it is unknown how well it performs in Bangladesh because no studies have validated the reliability of GWS data for the entire country. A few studies, that primarily used GWL data, have evaluated groundwater changes in Bangladesh.\(^15\)–\(^19\) Very few studies used GWS data obtained from Gravity Recovery and Climate Experiment (GRACE) and GLDAS data.\(^17\),\(^18\) However, most investigations were primarily focused on the northwestern section of the country, making them unable to estimate the overall groundwater
Accurate GWL prediction is critical for the long-term exploitation and management of crucial groundwater resources. Hydrogeological processes vary greatly in time and space; therefore, modeling the fluctuations in groundwater is challenging. Modeling groundwater flow in different hydrogeological settings may be done using both conceptual and process-based techniques. Groundwater change simulations using process-based models have prohibitively needs large amount of data. Despite tremendous effort and financial expenditures, the forecast accuracy of distributed numerical flow models has still not improved considerably in GWL prediction. Therefore, a dynamically predictive model which can account for the persistent trend and time-variant behaviour of hydrological variables is preferred for sustainable water resource planning and management. Empirical models such as MLR and ANN are attractive options in these situations since they provide relevant findings with fewer data points, require less effort, and are more affordable. MLR models have been employed in many hydrological studies despite their inability to address the model's input-output non-linearity. This is because the results are simple to apply, and the relationship between parameters is easier to grasp. Machine learning models like ANN have been used in recent years to overcome the drawback of MLR in mapping the nonlinear input-output relationship. Numerous studies have demonstrated that ANN can estimate GWL quite precisely. Rajaee reviewed many other studies on modelling GWL using ANN. The ANN technique is advantageous over traditional methods due to its suitability for modeling nonlinear and dynamic systems. It also does not necessitate the explicit mathematical explanation of the compound behavior of the processes.

In this study, spatial mean, variability, and trends of GLDAS GWS and BWDB GWL data for 2003-2019 were initially compared to evaluate the performance of GLDAS GWS in Bangladesh. MLR and ANN were then established using freely available rainfall, temperature, elevation, GWS, VTWSC, population, and irrigation data to predict GWL at any location at any point in time with higher accuracy and reliability. This is the first study to model the relationship between in-situ GWL data and openly available data to predict GWL for easy estimation of GWL. The model developed in this study can help the sustainable management of groundwater resources in the country.
2 Study Area

Bangladesh, situated close to the Bay of Bengal's mouth (Fig. 1), covers a sizable portion of the Bengal Delta created by the Ganges-Brahmaputra-Meghna (GBM) river system depositing sediment. The geographical coordinates of the country are as follows: 20°34'N, 26°38'N, 88°01'E, 92°41'E. The land is characterized by its many rivers and relatively flat terrain, which accounts for around 80% of the country's total land area. Nearly 70.1% of the land is used for agricultural purposes, 11.1% covers forest, and 18.8% is used for other purposes. The total population was 163 millions in 2019.

Bangladesh is the lowermost part of GBM, which drains 1.72 million km² of land. Essentially, just 8% of the watershed is occupied by Bangladesh. Surface water infiltration from external sources is essential for a part of the groundwater. During the Pleistocene and Holocene period, large rivers deposited sediments in the GBM delta, therefore constructing the delta and aquifer systems. A large number of aquifers are composed of Miocene to modern-day semi-consolidated to unconsolidated fluvial-deltaic sediments. Only the Holocene and Plio-Pleistocene Dupi Tila Sandstone formations are shallow enough to allow for groundwater collection. The average depth of the aquifers is between 30 and 130 metres. The principal water-bearing zone begins at a depth of 250-350 metres and is composed mostly of fine to very fine sand, which is occasionally interbedded with clay lenses and is typically interbedded with a silty clay bed.

Bangladesh, which straddles the Tropic of Cancer, has a humid tropical/subtropical climate with fairly high humidity, mild temperatures, and a substantial seasonal fluctuation in rainfall. According to the Köppen-Geiger climate classification, there are three distinct climates in the country: a tropical savannah climate, a tropical monsoon climate, and a temperate dry winter and hot summer climate (Aw; Fig. 1). The average annual rainfall is 2300 mm, with a range of 1,500 mm in the west to nearly 4,000 mm in the northeast. The average temperature is low in the winter (18°C), while high in the summer (30°C). Climate models indicate that Bangladesh's average temperature will rise by 2.4 °C and its annual rainfall would increase by 9.7 % by the end of the 21st century.
Figure 1. Location of groundwater monitoring wells over the climatic zone map of Bangladesh.\textsuperscript{46}
3 Methodology and Materials

3.1 Data & Sources

The weekly GWL data of the selected (based on the hydrographs) wells were collected from BWDB. The data request was made online through BWDB wet portal (http://www.hydrology.bwdb.gov.bd/).

GLDAS-2 CLSM's daily product has 33 parameters. In this study, GLDAS 2 CLSM data have been used that simulate shallow groundwater. GLDAS GWS daily data of 0.25° × 0.25° resolution from 2003-2020 (GLDAS 2.0) were downloaded from its website.

Selecting relevant (significant) input variables is a crucial part of developing MLR and ANN models. Some input variables are not valuable because they are redundant, noisy, or have no influence on the target output. This study employed GWS from GLDAS, elevation, temperature, rainfall from Asian Precipitation-Highly Resolved Observational Data Integration Towards Evaluation of the Water Resources (APHRODITE), population data from LandScan, irrigation data from FAO AQUASTAT, and variability in terrestrial water storage changes (VTWSC) estimated from Gravity Recovery and Climate Experiment (GRACE) by calculating the standard deviation of monthly terrestrial water storage changes (TWSC) data of that year.

Table 1 provides other relevant information of the data employed in the study.

GRACE was launched in March 2002 to estimate variations in Earth's field of gravity. By translating recorded gravity anomalies into changes in corresponding changes in water column height, the GRACE twin satellites monitor monthly variations in TWS. The monthly mass grid developed by the University of Texas at Austin's Center for Space Research (CSR) using RL05 spherical harmonic coefficients was employed in this investigation.

APHRODITE provides daily gridded precipitation and temperature dataset covering more than 50 years. They also offer elevation information. To model the water table, many writers have used APHRODITE climatic data as input parameters: precipitation and precipitation and temperature.

LandScan is the highest resolution worldwide population distribution data available (average over 24 hours), with a resolution of around 1 km (30° × 30°). LandScan data distribution models are tailored to each country and region's data conditions and geographical makeup.

The FAO AQUASTAT irrigated area (version 5) map depicts the area equipped for irrigation as a proportion of the total area of each cell of 10 × 10 km. This study used ArcGIS Desktop 10.8 and R programming software for data processing and analysis.
<table>
<thead>
<tr>
<th>Product</th>
<th>Resolution</th>
<th>Period</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TWSC</td>
<td>0.5° × 0.5°</td>
<td>Monthly</td>
<td>2002-2019</td>
</tr>
<tr>
<td>GWS</td>
<td>0.25° × 0.25°</td>
<td>Daily</td>
<td>2003-2019</td>
</tr>
<tr>
<td>Rainfall</td>
<td>0.25° × 0.25°</td>
<td>Daily</td>
<td>2003-2015</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.25° × 0.25°</td>
<td>Daily</td>
<td>2003-2015</td>
</tr>
<tr>
<td>Population</td>
<td>30” × 30”</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Description of the data employed in the study

3.2 Methods

This study adopted the subsequent measures to fulfill the goals of this study: (a) collection and extraction of data, (b) processing of all the data and bringing into the same format (0.25° × 0.25° spatial resolution), (c) generation of yearly time series of GWS and GWL, and calculation of yearly trends using nonparametric Sen slope estimator and modified M-K test, (d) generation of inputs for GWL prediction models from the data listed in Table 1, (e) development of MLR and ANN models to estimate mean GWL for a grid from the generated inputs, (f) employ ANN model to simulate mean GWL where in-situ GWL data are not available for Bangladesh.

3.2.1 Data processing

BWDB GWL data was examined thoroughly to determine its quality. The years 2003-2019 were chosen for the analysis. The data from each site was organized into a time series, histogram, and boxplot to visualize the gaps, outliers, and other irregularities. The samples with more than 1% of outliers, or the data contains sudden shifts and unpredictable fluctuations were thrown out. In the end, stations where the percentage of missing data was 10% or less was chosen. This method determined that 844 monitoring wells provided worthy data for analysis. In order to complete the groundwater hydrographs, this research used cubic spline interpolation (CSI). Several statistical methods, such as temporal interpolation with kriging, an autoregressive-moving-average model with exogenous inputs, multi-level regression, and
a seasonal autoregressive integrated moving average $^{64}$, have been developed for the imputation
of groundwater level data. GWL in Bangladesh often undergoes rapid shifts in a very short
amount of time. Therefore, for accurate imputation of missing GWL data in a series $^{65,66}$, every
tiny part of the series should be addressed independently.

Fig. 1 shows the placement of the observation wells. According to the location of the
wells, they are spread throughout the whole of Bangladesh with the exception of certain hilly
and forested areas in the southeast. Afterwards, the GWL well data were converted into $0.25^\circ \times 0.25^\circ$ spatial resolution grids, like GWS data. For this, the GWL of all wells within a grid
area was averaged (Fig. 4(c)). For other selected input data in MLR and ANN models, all the
data with any other spatial resolution were temporally averaged (except elevation) and
converted into $0.25^\circ \times 0.25^\circ$ resolution (Fig. 2). As BWDB groundwater observation wells are
not available for comparison and model development, only the grids were selected where GWL
data are available.
Figure 2. Continued.
Figure 2. Processed and gridded input data: a) mean temperature, b) mean rainfall, c) population, d) irrigation, e) elevation, and f) VTWSC. Along with these, GWS, as shown in Fig. 4(a), were used as inputs for model development.

3.2.2 Trend analysis

It is possible to identify and/or quantify trends using a variety of parametric, nonparametric, and mixed approaches. Due to its resiliency to missing and tied data, serial dependence, non-linearity, non-normality, and seasonality, nonparametric approaches are often used in climatic and hydrological research. The nonparametric modified M-K test and Sen slope estimator were employed in the present study. The Sen slope estimator is mainly employed to quantify changes per unit time while the modified M-K test evaluates the change significance by omitting the influence of autocorrelations.

3.2.3 MLR model

MLR method provides a concise explanation of the phenomenon being studied by employing a linear equation to characterise the link between a set of independent (explanatory) elements and a response (dependent) variable. The MLR’s basic structure is as follows:

\[ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i \]  

(1)
where $Y_i$ represents the $i^{th}$ observation of the dependent variable $Y$; $X_{1,i}, X_{2,i}, \ldots, X_{k,i}$ represent the $i^{th}$ observations of independent variables $X_1, X_2, \ldots, X_k$, respectively; $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are fixed but unknown parameters; and $e_i$ is a random variable that is normally distributed. MLR estimates the unknown parameters ($\beta_0, \beta_1, \beta_2, \ldots, \beta_k$) of an MLR model (Eq. 1). Consequently, the least-square technique's practical interpretation of the statistical regression model is given as:

$$Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \ldots + b_k X_{k,i} + e_i \quad (2)$$

where $i=1, 2, \ldots, n$; $b_0, b_1, b_2, \ldots, b_k$ represent the unstandardized regression coefficients of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ respectively; and $e_i$ is the residual for the $i^{th}$ observation.

### 3.2.4 Artificial Neural Network

ANN comprises nodes, often known as 'neurons,' as the network's processing units. Each neuron accepts and analyses input, and converts it to an output signal after being connected to other neurons. ANN is designed to perform 'biological neurons' functions artificially. As a result, creating ANN (or ANN models) involves five fundamental steps: (1) choosing influential inputs; (2) choosing an appropriate ANN topology; (3) construction the network; (4) training and validate the model; and (5) assessing the skill of the model. The procedure employed for developing ANN model for predicting GWL is shown using a flowchart (see Supplementary Fig. S1 online). Different phases of ANN are discussed briefly in the sections below.

This study employed a multilayer feedforward network (MLF) with a single hidden layer was chosen since it has been frequently utilized in groundwater simulation/prediction due to its ease of construction. The nodes in an MLF are often organized in layers, with one or multiple hidden layers between the input and output levels. Fig. 3 depicts a three-layer feedforward ANN with an input layer with 7 neurons, a hidden layer, and an output layer, as employed in this study. It has input nodes ($x_1, x_2, x_3, \ldots, x_6$) with bias ($B_{HK}$), hidden nodes ($h_1, h_2, h_3, \ldots, h_n$) with bias ($B_0$), and the output node ($Y_1$), $n$ denote the number of nodes in every layer (Fig. 3). The $W_{ik}$ denotes the interconnection weights between the input and hidden nodes, while $W_{ik}$ denotes the connection weights between the hidden and output nodes.
The prediction of the ANN model ($Y_f$), which is GWL in this case, can be expressed as follows:

$$\text{GWL} = f_T [B_0 + \sum_{k=1}^n \{W_k f_T(B_{Hk} + \sum_{i=0}^m W_{ik} P_i)\}]$$

where $B_0$ denotes the output layer bias (which consists of a single neuron); $W_k$ is the connection weight between the $k^{th}$ neuron in the hidden layer and the single neuron in the output layer. The connection weight between the input variable $i$ ($i = 1, m$) and the neuron $k$ of the hidden layer is designated as $W_{ik}$. $B_{Hk}$ is the bias at neuron $k$ ($k = 1, n$) of the hidden layer. The transfer function is denoted by $f_T(\theta)$, while $P_i$ designates the input parameter.

**Figure 3.** The multilayer feedforward network design utilized in the study (parameters described in the text) layer.

The network must first be trained after the ANN architecture has been chosen. The training aims to confirm the ANN captures the essential properties or input data patterns. The
model’s weights and biases are calculated during the training process. Backpropagation training, a two-step algorithm, is the most prevalent training algorithm of feedforward ANN. The input data is first transmitted forward in the first step to calculate the outputs. Then a backward step is conducted to alter the weight vectors between the layers to reduce model error. There are a variety of weight-optimization strategies for training ANNs. Many of which have been tested in this study (Table 4), the Levenberg-Marquardt (LM) algorithm was prioritized because of its ability to achieve rapid convergence.

The way in which a node responds to the complete input signal is set by its activation (transfer) function. Linear, hyperbolic tangent sigmoid, and logistic sigmoid activation functions are the most often used. The best-suited function was employed in the final ANN model, which considers the inclusion of the functions in both the hidden and output layers.

Hyperbolic tangent sigmoid function: \( \varphi(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} \) \hspace{1cm} (4)

Logistic sigmoid function: \( \varphi(v) = \frac{1}{1 + e^{-v}} \) \hspace{1cm} (5)

Linear function: \( \varphi(v) = v \) \hspace{1cm} (6)

where \( v \) in Eq. 4 stands for the weighted sum of inputs at the output layer, \( v \) in Eqs. 5 and 6 represent the weighted total for a node in the hidden layer, and \( e \) represents the natural exponential function.

The task of hidden-layer neurons is to establish links between a network's inputs and outputs. If case of insufficient hidden neurons, convergence could be challenging to accomplish during training. On the other side, if the network uses too many hidden neurons, it might lose its capability of generalization. Some researchers have shown that an optimization algorithm can be helpful in optimizing hidden neurons. However, the commonly used trial-and-error method was employed in this study.

Three aspects are relevant to weight optimization during training: the initial weight matrix, the error function, and the termination criteria, such as establishing the epoch number, specifying a target error objective, and fixing the least performance gradient. Eq. 7 was then used to optimize these weights. For weight optimization, the following objective function (E) was utilized:

\[ E = \frac{1}{2n} \sum_{i=1}^{n} (h_{oi} - h_{pi})^2 \] \hspace{1cm} (7)
where \( h_{oi} \) represents the observed groundwater level at \( i^{th} \) time, \( h_{pi} \) represents the anticipated groundwater level at \( i^{th} \) time, and \( n \) represents the total number of observations.

MATLAB was used to complete the ANN modeling process, where 70% of the data were taken for training, 15% for validation, and 15% for testing. This proportion was selected after testing different proportions to see how well the ANN model performed.

### 3.2.5 Model's Performance Assessment

The efficiency of the developed MLR model was determined using a set of conventional statistical indicators, which included: multiple correlation coefficient (R), coefficient of multiple determination (R\(^2\)), adjusted R\(^2\), standardized regression coefficient (\( \beta_j \)), standard error (SE), standard error of estimate (SEE), t-test, F-test, and p-test, etc. And for ANN models two statistical metrics, namely mean square error (MSE) and R, were used to assess their performance.

### 4 Results

#### 4.1 Comparison between GWL and GWS data

For validating GWS data for Bangladesh, the spatial mean, variability, and trends in GWS were compared to that for GWL. The mean GWS in Bangladesh, as shown in Fig. 4(a), ranged from 0.63 m to 1.74 m thick water column. The highest GWS, from 1.21 to 1.74 m, was found in the country's northeastern region. Higher values were recorded in the centre east to southeast areas, which fall in tropical monsoon zone (Am; Fig. 1). The lowest values, ranging from 0.63 to 0.72 m, were observed in the northwest to central west areas, mostly in temperate, dry winter, hot summer (Cwa) and tropical savannah (Aw) zones, but there is a southeast strip in tropical monsoon zone. Figure 4(c) shows that the average GWL was between 0.92 and 26.35 m below ground level. Central and northern parts have the lowest GWL, measuring between 11.40 and 26.35 m below ground level. The southern region, as well as parts of the northeast and far northwest, had the highest GWL levels (0.92 to 2.99 m).

Fig. 4(b) and 4(d) show the spatial pattern in the trends of GWS and GWL, respectively. The rate of change is presented using different colors on the maps. The significant change at the 95% level of confidence is also depicted in the figure. The results showed a decreasing trend in GWS throughout the country. The decreases were highest in the northwest by -0.013 to -0.017 m/year and the lowest in the northeast and south by 0 to -0.002 m/year. Fig. 4(d) also shows a declination in GWL throughout the country. The highest rate of declination was in the
central and northwest parts by -0.120 to -0.272 m/year. A mild increase was also witnessed in many places throughout the country, especially in the south.

The monthly variability in GWS and GWL for each year was calculated to produce GWS and GWL variability time series at each location from 2003 to 2019. This series was employed to calculate the average GWS and GWL variability and their changes over time at each grid. The goal was to illustrate the seasonal and yearly fluctuations in GWS and GWL. GWS variability was found to be greater in the southeast and northwest and lower in the northeast and southwest. GWL variability, on the other hand, was greatest in the northwest and north-central areas (Fig. 5(a)). In such regions, the GWL ranged as high as 3.95 m.

The trends in GWS variability showed increased variability in the south and decreased variability in the country's north. In contrast, the trends in GWL variability showed not such a spatial pattern (Fig. 5(b)). The highest increase was in the central areas of the country. However, there were a few localised areas where GWL variability was lower than elsewhere. The wide range in GWL variability suggests that GWL deviation from the mean throughout much of Bangladesh. Also, the scatterplot in Supplementary Fig. S2 online shows a weak, negative, and linear relationship between mean GWS and GWL at 183 grids covering Bangladesh with an R of 0.15.
**Figure 4.** Spatial distribution of mean GWS (a) and GWL (c) in m; the rate of change in GWS (b) and GWL (d) in m/year, for 1995–2019 using Sen's slope method. The color ramp.
represents change per year, whereas the white dot sign in the map represents the significant decrease or increase in GWL derived using the M-K test.

Figure 5. Same as Fig. 4, but for GWL variability within a year
4.2 MLR Model

The results obtained in prediction GWL using the MLR model are summarized in Table 2, while the overall performance of MLR is shown in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>p</th>
<th>R</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-9.515</td>
<td>-1.51</td>
<td>.134</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWS</td>
<td>1.752</td>
<td>.114</td>
<td>1.10</td>
<td>.273</td>
<td>-.151</td>
<td>.083</td>
</tr>
<tr>
<td>Elevation</td>
<td>.013</td>
<td>.126</td>
<td>1.57</td>
<td>.117</td>
<td>-.053</td>
<td>.118</td>
</tr>
<tr>
<td>Rainfall</td>
<td>-.001</td>
<td>-.329</td>
<td>-3.18</td>
<td>.002</td>
<td>-.312</td>
<td>-.233</td>
</tr>
<tr>
<td>Population</td>
<td>8.175E-006</td>
<td>.337</td>
<td>5.21</td>
<td>.000</td>
<td>.370</td>
<td>.365</td>
</tr>
<tr>
<td>Temperature</td>
<td>.423</td>
<td>.144</td>
<td>1.83</td>
<td>.050</td>
<td>.022</td>
<td>.137</td>
</tr>
<tr>
<td>Irrigation</td>
<td>.006</td>
<td>.088</td>
<td>0.73</td>
<td>.463</td>
<td>.349</td>
<td>.055</td>
</tr>
<tr>
<td>VTWSC</td>
<td>.028</td>
<td>.219</td>
<td>2.03</td>
<td>.044</td>
<td>.292</td>
<td>.151</td>
</tr>
</tbody>
</table>

Table 2. MLR model results listing the coefficients, the goodness of fit parameters, and collinearity statistics

<table>
<thead>
<tr>
<th>R</th>
<th>R²</th>
<th>Adjusted R²</th>
<th>SEE</th>
<th>F-statistic</th>
<th>p</th>
<th>Durbin-Watson statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>.549</td>
<td>.301</td>
<td>.273</td>
<td>2.1586609</td>
<td>10.833</td>
<td>0.000</td>
<td>1.440</td>
</tr>
</tbody>
</table>

Table 3. Summary of MLR model's overall performance

In Table 3, the 'β_k' (standardized regression coefficients) indicates each independent variable's relative contribution to the GWL prediction. The regression equation's weights or raw or unstandardized regression coefficients are referred to as 'b_k'. Consequently, the MLR model created using the regression approach to predict GWL is:

\[ GWL = -9.515 + 1.752 \times \text{GWS} + 0.013 \times \text{Elevation} - 0.001 \times \text{Rainfall} + 8.175E-006 \times \text{Population} + 0.423 \times \text{Temperature} + 0.006 \times \text{Irrigation} + 0.028 \times \text{VTWSC} \] (8)

It is obvious from Table 3 that population \((t = 5.21, p = 0.000)\), rainfall \((t = -3.18, p = 0.002)\), VTWSC \((t = 2.03, p = 0.044)\), temperature \((t = 5.16, p = 0.050)\) are significant explanatory variables, indicating higher t values and very low p-level (i.e., 0.0000). In contrast,
elevation \((t = 1.57, p = 0.117)\), GWS \((t = -2.62, p = 0.001)\), and irrigation \((t = -2.59, p = 0.001)\) are not significant in predicting GWL. However, the elevation, irrigation, and GWS were included in Eq. 8 as they can have a combined influence in predicting GWL. \(SE\) and \(R\) (both zero-order and partial) follow a similar pattern. In this context, the correlation between the dependent and independent variables is referred to as zero-order correlation because no other variables' effects were considered. In partial correlation, the impact of other controlling variables is removed. The variance inflation factor (VIF) and tolerance were used to evaluate the multicollinearity to what extent it existed. As a general rule, strong multicollinearity is indicated by a variable with a VIF value larger than 5 or a tolerance lower than 0.2. Table 2 shows that for all independent variables included in creating the MLR models, the value of VIF is much less than 5 and the tolerance is more than 0.2. Therefore, there is no multicollinearity in the chosen independent variables.

Additionally, the MLR model's overall significance was examined using the F-test. It is evident from the p-value \((p = 0.000)\) F-statistic \((F = 10.833)\) that the model is significant, meaning that the independent variables adequately account for the variability in GWL fluctuations. However, \(R\), multiple \(R^2\), and adjusted \(R^2\) indicate the significance of the model to be moderate. The Durbin-Watson statistic was used to examine the autocorrelation of residuals. At a 95% confidence level, the residuals' Durbin-Watson statistic was 1.44. The result suggests statistical evidence that the error terms are positively autocorrelated at lag 1 in the time series of GWL.

### 4.3 ANN Model

Different combinations of networks, training, transfer, adaptation learning, and performance functions were employed for ANN model training to find the best combination for hyperparameters. The different number of hidden neurons for each combination was also tried. Table 4 shows the result obtained from different training functions. The other functions and network types are presented in Table 5. Table 4 shows the best results using the Levenberg-Marquardt training function \((R = 0.95, \text{MSE} = 0.64)\), followed by Bayesian Regularization \((R = 0.93, \text{MSE} = 0.71)\), Scaled Conjugate Gradient \((R = 0.92, \text{MSE} = 0.74)\), Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi-Newton \((R = 0.90, \text{MSE} = 0.75)\), and other functions. The best result was obtained using the combination of feedforward backpropagation network, Levenberg-Marquardt training function, gradient descent with momentum adaption learning.
function, hyperbolic tangent sigmoid transfer function, and mean squared error (MSE) performance function, as presented in Table 5.

<table>
<thead>
<tr>
<th>Training Function</th>
<th>Number of hidden neurons</th>
<th>R</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable Learning Rate Gradient Descent</td>
<td>13</td>
<td>0.67</td>
<td>3.61</td>
</tr>
<tr>
<td>Conjugate Gradient with Powell</td>
<td>9</td>
<td>0.72</td>
<td>3.08</td>
</tr>
<tr>
<td>Resilient Backpropagation</td>
<td>8</td>
<td>0.73</td>
<td>2.99</td>
</tr>
<tr>
<td>Polak-Ribiére Conjugate Gradient</td>
<td>11</td>
<td>0.74</td>
<td>2.92</td>
</tr>
<tr>
<td>Fletcher-Powell Conjugate Gradient</td>
<td>12</td>
<td>0.76</td>
<td>2.66</td>
</tr>
<tr>
<td>One Step Secant</td>
<td>12</td>
<td>0.76</td>
<td>2.71</td>
</tr>
<tr>
<td>BFGS Quasi-Newton</td>
<td>9</td>
<td>0.9</td>
<td>0.75</td>
</tr>
<tr>
<td>Scaled Conjugate Gradient</td>
<td>10</td>
<td>0.92</td>
<td>0.74</td>
</tr>
<tr>
<td>Bayesian Regularization</td>
<td>9</td>
<td>0.93</td>
<td>0.71</td>
</tr>
<tr>
<td>Levenberg-Marquardt</td>
<td>10</td>
<td>0.95</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Table 4 Performance of different training functions

Table 5 Specifications for final ANN model

Table 5 reveals the overall R-value of 0.95 (0.95 for training, 0.96 for validation, and 0.90 for testing; Fig. 6) and an MSE value of 0.64. The weights and biases generated from this training are given in Table 6, which can be used to regenerate the network for GWL prediction.
Fig. 6. Performance of ANN model

Fig. 7 shows the relative importance of different input variables in ANN training. The results indicate population as the highest influencing input in ANN training, followed by rainfall, VTWSC, elevation, GWS, temperature, and irrigation.

<table>
<thead>
<tr>
<th>Hidden neuron (HN)</th>
<th>Weight</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>no.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HN 1 (k=1)</td>
<td>3.9359</td>
<td>1.4888</td>
</tr>
<tr>
<td></td>
<td>-2.3393</td>
<td>-2.1453</td>
</tr>
<tr>
<td></td>
<td>-1.5527</td>
<td>-5.3105</td>
</tr>
<tr>
<td></td>
<td>-5.41843</td>
<td>2.545</td>
</tr>
<tr>
<td></td>
<td>-5.6518</td>
<td>0.48179</td>
</tr>
<tr>
<td>HN 2 (k=2)</td>
<td>-0.10975</td>
<td>0.31356</td>
</tr>
<tr>
<td></td>
<td>-0.96679</td>
<td>-1.1747</td>
</tr>
<tr>
<td></td>
<td>1.7922</td>
<td>-1.6637</td>
</tr>
<tr>
<td></td>
<td>-1.6056</td>
<td>-3.7426</td>
</tr>
<tr>
<td></td>
<td>-2.0943</td>
<td></td>
</tr>
<tr>
<td>HN 3 (k=3)</td>
<td>1.2306</td>
<td>2.3356</td>
</tr>
<tr>
<td></td>
<td>2.0073</td>
<td>-2.7986</td>
</tr>
<tr>
<td></td>
<td>0.097563</td>
<td>-4.7154</td>
</tr>
<tr>
<td></td>
<td>0.94473</td>
<td>0.38314</td>
</tr>
<tr>
<td></td>
<td>3.0659</td>
<td></td>
</tr>
<tr>
<td>HN4 (k=4)</td>
<td>-0.2341</td>
<td>-0.3006</td>
</tr>
<tr>
<td></td>
<td>1.4528</td>
<td>0.10213</td>
</tr>
<tr>
<td></td>
<td>1.5642</td>
<td>-1.0243</td>
</tr>
<tr>
<td></td>
<td>0.62849</td>
<td>3.6601</td>
</tr>
<tr>
<td></td>
<td>-0.29346</td>
<td></td>
</tr>
</tbody>
</table>
Table 6. Connection weights and biases of final ANN model

| HN 5 (k=5) | -4.0517 | -4.3297 | 5.2031 | 0.70048 | 0.90006 | 2.3688 | 3.1789 | -0.52562 | 0.42161 |
| HN 6 (k=6) | -3.3457 | 1.0343 | 1.2057 | 1.3426 | 0.93627 | 0.80308 | 0.22901 | -0.42983 | -0.56644 |
| HN 7 (k=7) | -0.69892 | -4.8222 | -0.13239 | 2.8573 | -0.79932 | 4.6237 | 2.7112 | 0.57117 | -3.0504 |
| HN 8 (k=8) | -3.3395 | 1.8932 | -0.46469 | -0.8387 | 0.46286 | -0.2832 | -1.1937 | 0.85997 | -0.16256 |
| HN 9 (k=9) | 4.8535 | 2.1274 | 0.75186 | 2.4445 | -1.9791 | -7.1745 | 2.0173 | -0.2435 | -0.07174 |
| HN 10 (k=10) | 1.5314 | 3.9166 | 0.91078 | 2.0042 | -0.7953 | 1.7354 | -2.6909 | 0.40058 | 3.0986 |

The line graph and grid maps in Fig. 8 and Supplementary Fig. S3 online show the comparison between actual and ANN-generated mean GWL at each grid. The result showed almost identical GWL indicating high accuracy of the trained ANN model.
Discussion

The objective of this study has been to either identify a reliable and easily obtainable data source for regular observation of groundwater or establish a model that can accurately predict the level of groundwater from freely available data. For that purpose, GLDAS GWS was compared with the BWDB GWL data. The results indicate insignificant spatial association in the mean, variability, and trends (Fig. 4, 5, and Supplementary Fig. S2 online) between GWS and GWL in Bangladesh. As a result, relying solely on GWS data for regular groundwater monitoring would be insufficient. Consequently, modeling GWL becomes a more viable and necessary alternative.

In modeling MLR and ANN where various assemblages of performance functions, networks, training, transfer, and adaptation functions were performed. The results showed that the ANN model’s skill was considerably better (with an overall R of 0.95 against 0.54) in predicting GWL using the selected input variables. Contrary to the ANN approach, the MLR technique is easier to adopt, less labor- and time-intensive, and necessitates less skill. However, despite these significant practical benefits, the accuracy of the MLR technique was not up to the mark in simulating such a nonlinear and dynamic system. As a result, it is recommended to

Figure 8. Map showing the comparison between actual mean GWL and ANN model generated mean GWL.
employ ANN models rather than MLR in predicting GWL. However, there were some similarities in the ranking of significance or importance of the input variables; the only difference was that in ANN, GWS was more significant than temperature, whereas, in MLR, the opposite was true.

There are some places in Bangladesh without any BWDB groundwater observation wells. Despite the fact that the developed ANN model was validated and tested using just 30% of the data (Fig. 6), the developed ANN model was utilized to predict GWL in those areas to demonstrate its usefulness (see Supplementary Fig. S4 online). The results were coherent with the adjacent actual mean GWL. But significantly higher values were found in the hilly southeastern regions (Chittagong Hill Tracts). The primary reason behind that is the GWL indicates the depth of the GWL below the ground surface, and the elevation in hilly areas is naturally higher than the persistent GWL. Different studies have also found the GWL in hilly areas is significantly low.

This study focused on the spatial analysis of the average values of the employed parameters rather than taking into account any time step in the model development. This was due to the lack of coherence in the temporal resolution and the time frame of available data.

**6 Conclusion**

This study developed a method for easy and regular monitoring of groundwater status in Bangladesh. Initially, spatial mean, variability, and trends of GLDAS GWS and BWDB GWL data for 2003-2019 were calculated and compared. MLR and ANN models were then established using freely available rainfall, temperature, elevation, GWS, VTWSC, population, and irrigation data to predict GWL data for any location. The key findings from the study can be summed up as follows. (1) Reliable monitoring of groundwater status using only GLDAS GWS data is insufficient as its spatial mean, variability, and trends have a weak correlation with in-situ GWL data. (2) MLR and ANN models can be used to predict GWL for any location in Bangladesh with acceptable accuracy. (3) ANN model showed remarkably higher accuracy than MLR. Therefore, ANN is recommended for predicting GWL with greater accuracy.

The findings will aid practitioners in conducting frequent GWL level monitoring to achieve sustainable and balanced groundwater management to conserve and regenerate the country's groundwater system. The ANN technique can be used as a cost-effective option for GWL simulation/prediction in situations when professional competence and time are limited; however, field data are still favorable (i.e., noise-free and of good quality). The methodology given in this study is easily transferable to other parts of the world, regardless of
hydrogeological conditions. As a result, the technique and conclusions are extremely beneficial to both developing and developed country researchers working on groundwater management and conservation and practicing hydrogeologists.

Acknowledgments

The authors are grateful to the Bangladesh Water Development Board for providing groundwater table data. The authors are also grateful to Center Space research, University of Texas, Austin, GLDAS, APHRODITE, LandScan, and FAO AQUASTAT for providing the data through their web portal.

Conflict of interest

The authors declare no conflict of interest.

Funding

Not Applicable

Availability of data and materials

Climate data used in this study are available in the public domain. Requests can be made for groundwater table data online through the official website of Processing and Flood Forecasting Circle, BWDB (http://www.hydrology.bwdb.gov.bd/).

Code Availability

The codes used for data processing can be provided on request to the corresponding author.

Ethics approval

Not Applicable

Consent to participate

Not Applicable

Consent for publication

All the authors consented to publish the paper.
References


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplimentaryInfoManuscriptANNGW.docx