hydrology

MDPI

Article

A Comparison of Ensemble and Deep Learning Algorithms to Model Groundwater Levels in a Data-Scarce Aquifer of Southern Africa

Zaheed Gaffoor [1,2,*], Kevin Pietersen [3], Nebo Jovanovic [2], Antoine Bagula [4], Thokozani Kanyerere [2], Olasupo Ajayi [4] and Gift Wanangwa [5]

[1] IBM Research Africa, Johannesburg 2001, South Africa
[2] Department of Earth Sciences, University of the Western Cape, Cape Town 7535, South Africa; njovanovic@uwc.ac.za (N.J.); tkanyerere@uwc.ac.za (T.K.)
[3] Institute for Water Studies, University of the Western Cape, Cape Town 7535, South Africa; kpietersen@mweb.co.za
[4] Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa; abagula@uwc.ac.za (A.B.); ooajayi@uwc.ac.za (O.A.)
[5] Groundwater Division, Water Resources Department of the Ministry of Water and Sanitation, Tikwere House, Lilongwe 207203, Malawi; 4055140@myuwc.ac.za
* Correspondence: zaheed.gaffoor1@ibm.com; Tel.: +27-11-276-9200

Abstract: Machine learning and deep learning have demonstrated usefulness in modelling various groundwater phenomena. However, these techniques require large amounts of data to develop reliable models. In the Southern African Development Community, groundwater datasets are generally poorly developed. Hence, the question arises as to whether machine learning can be a reliable tool to support groundwater management in the data-scarce environments of Southern Africa. This study tests two machine learning algorithms, a gradient-boosted decision tree (GBDT) and a long short-term memory neural network (LSTM-NN), to model groundwater level (GWL) changes in the Shire Valley Alluvial Aquifer. Using data from two boreholes, Ngabu (sample size = 96) and Nsanje (sample size = 45), we model two predictive scenarios: (I) predicting the change in the current month's groundwater level, and (II) predicting the change in the following month's groundwater level. For the Ngabu borehole, GBDT achieved R^2 scores of 0.19 and 0.14, while LSTM achieved R^2 scores of 0.30 and 0.30, in experiments I and II, respectively. For the Nsanje borehole, GBDT achieved R^2 of -0.04 and -0.21, while LSTM achieved R^2 scores of 0.03 and -0.15, in experiments I and II, respectively. The results illustrate that LSTM performs better than the GBDT model, especially regarding slightly greater time series and extreme GWL changes. However, closer inspection reveals that where datasets are relatively small (e.g., Nsanje), the GBDT model may be more efficient, considering the cost required to tune, train, and test the LSTM model. Assessing the full spectrum of results, we concluded that these small sample sizes might not be sufficient to develop generalised and reliable machine learning models.

Keywords: groundwater levels; machine learning; gradient boosting decision trees; long short-term memory neural networks; Southern Africa

check for updates

Citation: Gaffoor, Z.; Pietersen, K.; Jovanovic, N.; Bagula, A.; Kanyerere, T.; Ajayi, O.; Wanangwa, G. A Comparison of Ensemble and Deep Learning Algorithms to Model Groundwater Levels in a Data-Scarce Aquifer of Southern Africa. *Hydrology* **2022**, *9*, 125. https://doi.org/10.3390/hydrology9070125

Academic Editor: Roohollah Noori

Received: 4 June 2022
Accepted: 6 July 2022
Published: 15 July 2022

1. Introduction

Data-driven tools and techniques have become a ground-breaking new paradigm in information discovery [1,2]. This is especially true for artificial intelligence and machine learning-based approaches. Machine learning, a subset of artificial intelligence-based techniques, has been shown to handle large and high dimensional data and model the non-linear processes better than traditional data-driven techniques [1]. Applications in various industries and scientific domains, such as medicine [3,4], astronomy [5], earth sciences [6–8], and commerce [9], have illustrated the effectiveness and potential of machine learning as a decision support tool.

While still relatively novel, machine learning has gained traction in hydrology and the hydrogeology domain [10–13]. Studies have demonstrated the use of machine learning for tasks such as groundwater level modelling [14–18], groundwater quality modelling [19,20], groundwater exploration [21], and groundwater storage forecasting [22]. Ahmadi and Tao [12,13] provide an in-depth review of the use of the machine learning algorithms for groundwater level modelling. Their results indicate classical machine learning approaches to be the most widely adopted set of algorithms. In fact, in a number of studies reported by [13] the models demonstrated good performance with relatively small sample sizes, further highlighting the potential of machine learning in groundwater modelling applications.

Deep learning techniques, on the other hand, are not as widely applied to groundwater-related problems, but are increasingly being adopted, having also shown promising results. Raheja [23] showed that deep neural networks outperformed the classical machine learning approaches in predicting groundwater quality. Huang [24] demonstrated that a Long-Short Term Memory Neural Network (LSTM-NN), a deep learning model, outperformed classical machine learning approaches in predicting groundwater recharge in an aquifer in South-Western Australia. In addition, Malakar [25] demonstrated better performance from an LSTM model than a feedforward neural network and conventional recurrent neural network in predicting groundwater levels in numerous boreholes across India. On the other hand, Wunsch [26] compared the performance of several advanced machine learning approaches, namely non-linear autoregressive networks with exogenous inputs (NARX), long short-term memory and convolutional neural networks (CNN), to forecast groundwater levels. Their results showed that while NARX performed better on average, the deep learning models were superior when previous time steps were included as an input feature. Thus, it is apparent that deep learning methods are quickly proving to be efficient and capable tools for groundwater modelling.

However, one of the drawbacks of data-driven approaches is that they invariably rely on large volumes of data to sufficiently train and test the algorithms, which is especially true for deep learning models. In general, larger, more homogenous datasets lend themselves to better accuracy in the performance of machine learning algorithms. In fact, according to Ahmadi [12], a minimum of 10 to 12 years of monthly groundwater level data is needed to develop machine learning algorithms of acceptable levels of skill. While this level of groundwater data may be relatively abundant in developed nations, the opposite is true in developing nations. For example, the Southern African Development Community (SADC) and its member states, which are heavily dependent on groundwater resources [27], tend to have poorly developed groundwater datasets. This results from limited monitoring and data collection schemes operating across the SADC [28,29]. Understanding the applicability of these techniques in these data-scarce conditions is important in realising their potential from a groundwater management perspective in SADC [30].

The use of machine learning techniques within SADC for groundwater modelling, has seen a slow adoption [12,13]. Several studies, such as [14,31–33], have explored machine learning for groundwater modelling in Southern Africa. Kombo [33] reported good skill using a hybrid K-Nearest Neighbour-Random Forest (KNN-RF) to predict groundwater level fluctuations in fractured rock aquifers in Kenya, while [31,32] demonstrated reasonable skill in predicting groundwater levels in karts aquifers in South Africa. Nonetheless, additional testing is needed to fully understand the applicability of these models in groundwater modelling applications, and especially within the context of data-scarce aquifers [13].

The purpose of this study is thus to compare the usefulness of two popular machine learning algorithms within a data-scarce and reasonably unexplored aquifer of SADC, the Shire Valley Alluvial aquifer (Figure 1). Specifically, Gradient Boosting Decision Tree (GBDT), a form of ensemble learning technique, and Long-Short Term Memory Neural Network (LSTM-NN), a form of deep learning technique, are applied to forecast groundwater level changes in individual boreholes (with limited and irregular data) across several temporal horizons. In addition, groundwater big data sources such as remote sensing data, data from land-surface models, and local in-situ data are integrated and utilised to train and test the models and provide insights into their predictive usefulness. This further highlights the role of remote sensing, and global earth system models to develop machine learning models, where in-situ groundwater level measurements alone fall short. Ultimately the results allude to the potential, or not, of these techniques to support groundwater management in data-scarce regions of SADC.

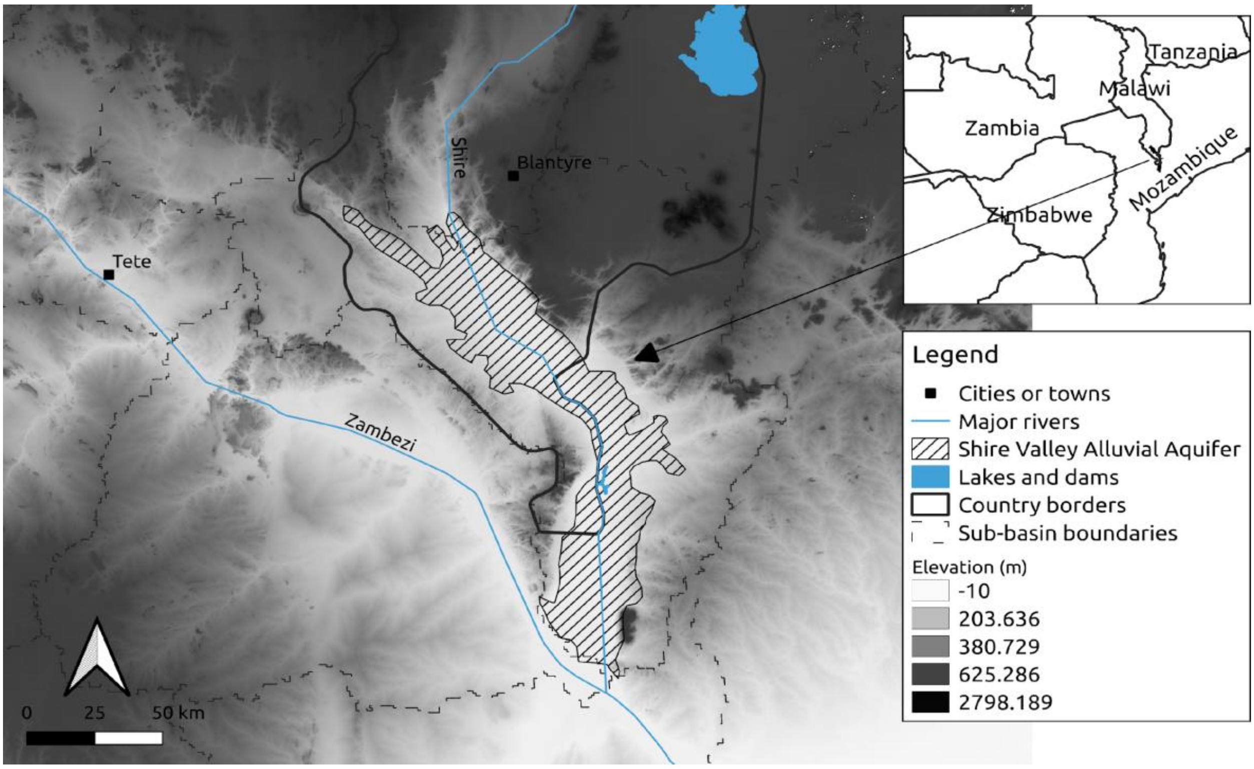

Figure 1. Map showing the location of the Shire Valley Alluvial Aquifer.

2. The Shire Valley Alluvial Aquifer System

The Shire Valley Alluvial Aquifer is a surficial alluvial transboundary aquifer between Malawi and Mozambique [34]. The aquifer lies within a sub-catchment of the Zambezi River basin and is drained by the Shire River, one of the main tributaries to the Zambezi River [35]. The Shire River originates as the main outflow from Lake Malawi to the north. The aquifer and overlying landscape are bordered by prominent escarpments such as the Thyolo escarpment to the north-east, and the highlands to the south-west [35], forming a typical valley landscape.

The alluvial aquifer is a vital freshwater resource extensively used for irrigation and domestic use by rural communities [36]. Abstraction occurs mainly through shallow boreholes and open wells, while in recent times, deeper boreholes have been developed. The aquifer suffers from many concerns, such as water quality degradation, the consequences of climate change and variability, poor institutional arrangements, and a lack of surface and subsurface data [36,37].

3. Methods and Data

To achieve the objective of this study, experiments were set up to train and test GBDT (Section 3.1.1) and LSTM-NN (Section 3.1.2) to model groundwater level changes in an aquifer considered to have limited long-term groundwater level data, the Shire Valley Alluvial Aquifer. These two machine learning algorithms have previously shown great skill in modelling groundwater levels [18,25,26,38,39], and invariably allow a comparison between ensemble and deep learning approaches. In addition, several studies have shown that these two algorithms outperformed other popular algorithms in their respective machine learning class [13,40,41]. Tree models in particular appear to demonstrate good skill with small sample sizes [13]. GBDT algorithms, and its component decision trees are connected sequential which helps minimize overall errors (improve classification/prediction accuracies). This is one of the primary reasons GBDT is considered as a better learner compared to other models. In addition, this work sought to model groundwater levels using historic data; however, the inherent autocorrelation between periodic groundwater levels may play a role in forecasting future conditions. LSTM-NN was thus selected for this work because of its innate ability to learn the dependency between data entries in a sequential dataset and use it to make future predictions about changes in groundwater levels.

The machine learning algorithms were trained and tested on individual data points (boreholes), with sufficient time-series data to warrant monthly groundwater level anomaly predictions. The data for each borehole was subjected to two different prediction scenarios when training and testing the machine learning algorithms: (I) predicting the groundwater level change for the current month, using hydroclimatic inputs for the current, hydroclimatic inputs of the previous month, and the groundwater level change of the previous month; (II) predicting the following month's groundwater level change, using hydroclimatic inputs for the current month, the groundwater level change of the current month, hydroclimatic inputs of the previous month, and the groundwater level change of the previous month. Finally, the analysis performed in this study was carried out in a python environment, using open-source libraries.

3.1. *Machine Learning Algorithms*

3.1.1. Gradient Boosting Decision Trees

Modern decision trees are non-parametric supervised statistical learning algorithms used as a predictive tool for numerical estimations. These classical machine learning approaches are excellent tools for regression, classification, and ranking tasks. Decision trees have several key advantages: easy to interpret, handling missing values, not being influenced by outliers, not needing a priori information, handling irrelevant features, requiring minimal data preparation, and not being affected by non-linear relationships [18,42]. Modern decision tree algorithms are also highly efficient and accurate [43].

Figure 2 presents a basic decision tree model illustrating the estimation of groundwater level changes using a set of input features. The basic structure includes nodes and branches. The root nodes represent a decision (typically a binary question such as yes or no) that needs to be made and that partitions the data space (predictor space) into two homogenous regions [44]. As we move further along the decision path, intermediate split nodes (or chance nodes) are encountered, representing another decision to split the data further. Finally, a leaf node is reached, representing the prediction made about a target (predictant) variable. The leaf node that is reached depends on the decisions made along the tree. The learning process during decision tree training adjusts the rules at split nodes as successive pairs of training examples are ingested into the model [45]. This ultimately results in a more accurate prediction of the target variable.

Technically speaking, decision trees split the data space into homogenous regions based on conditions (or decisions). The idea behind fitting a good decision tree is to ensure that the data space is split homogenously until the lead node is reached [42]. The training process ensures the splits represent the most homogenous partitioning of the data space, such that unbiased decisions (or predictions) are reached at leaf nodes. A number of metrics are used to measure node impurities, such as entropy, Gini index, classification error, information gain, and gain ratio, amongst others [46]. This helps grow the tree until the actual target value is reached.

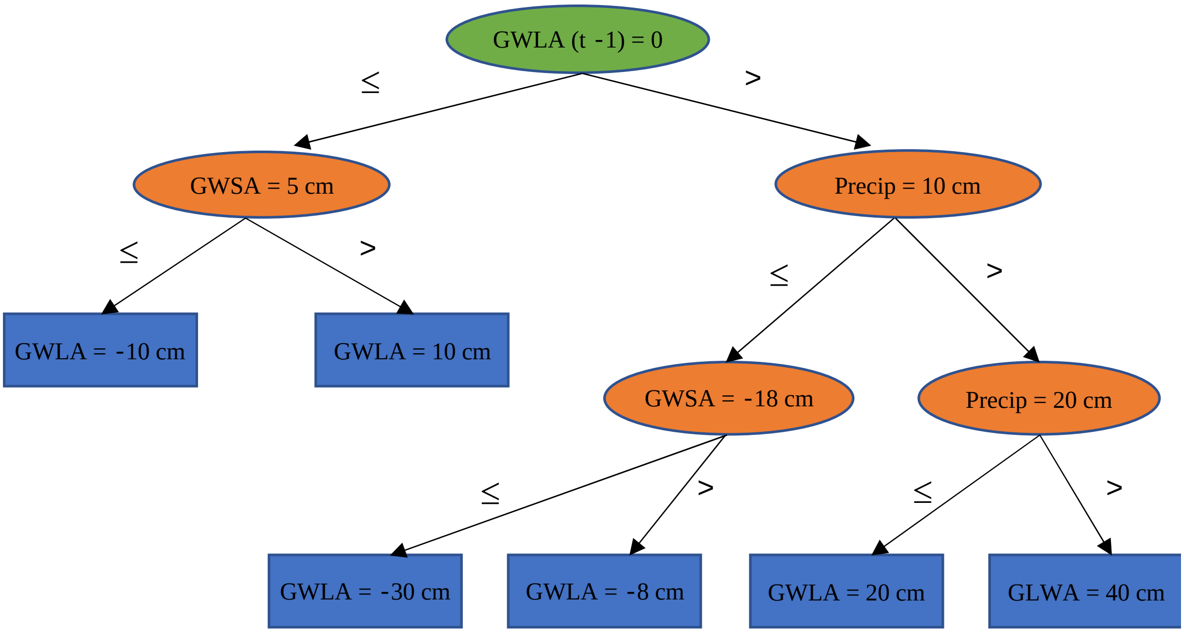

Figure 2. Conceptualised schematic of a traditional decision tree to determine the groundwater level change (GWLA). The root node is coloured green, intermediate split nodes are orange, and leaf nodes are blue. Data are routed through the split node based on rules/conditions until a leaf node (i.e., decision) is reached. GWLA (t − 1) = Previous groundwater level change, GWSA = Groundwater storage anomaly, Precip = Precipitation.

Gradient boosting is an ensemble technique (others include bagging, stacking, and model averaging) used in machine learning to combine a number of weak learners (decision trees) into a more robust model [44,45]. Instead of fitting one strong decision tree, gradient boosting relies on fitting successive weak learners to the same data. At each successive training iteration, the residual errors of the previous iteration are propagated forward to further reduce the gradient of a loss function as boosting progress [47].

In this study, a LightGBM GBDT model is used [43]. This model includes additional features, such as using a histogram-based approach to find node points in the predictor space for optimal splitting. The LightGBM implementation also includes the exclusive feature bundling (EFB) method, which reduces the complexity of building the histogram by grouping individual features into feature bundles. In addition, trees are grown leaf-wise instead of depth-wise. The framework can be implemented using the open-source python package, LightGBM (https://lightgbm.readthedocs.io/en/latest/, (accessed on 11 July 2022)).

3.1.2. Long Short-Term Memory (LSTM) Neural Networks

In sequencing learning problems, the input from previous states is important to the learning process, particularly in time-series data with inherent temporal auto-correlation, as previous conditions affect future signals. Initially, conventional recurrent neural networks (RNNs) were considered to handle the sequencing learning problems by storing information about past time steps. However, RNNs suffer from the vanishing gradient problem that significantly affects their ability to store long-term information, which is needed for time-series forecasting [48]. By combining LSTM with RNN, the Long short-term memory neural networks were born, a type of recurrent neural network within deep learning, specially designed to handle long order dependent sequences [49].

LSTM-NN was specifically designed to alleviate this problem by adding one or more connected memory cells in the LSTM block [50]. The memory cell allows past information to be stored longer than conventional RNN blocks. Also, memory gates (input, output and forget gates) are added to the LSTM block to control the flow of information into and out of the cells (Figure 3b). Typically, the memory block holds information about the input data and the previous steps. While the input and output gates control the information into and out of the memory block, the forget gate is designed to clear the old memory at a specific time, favouring more recent memory [51]. This functionality allows LSTM-NN networks to comprehend better and learn sequence-dependent prediction tasks. To simplify and avoid excessive acronyms, we refer to this hybrid LSTM-NN model as LSTM.

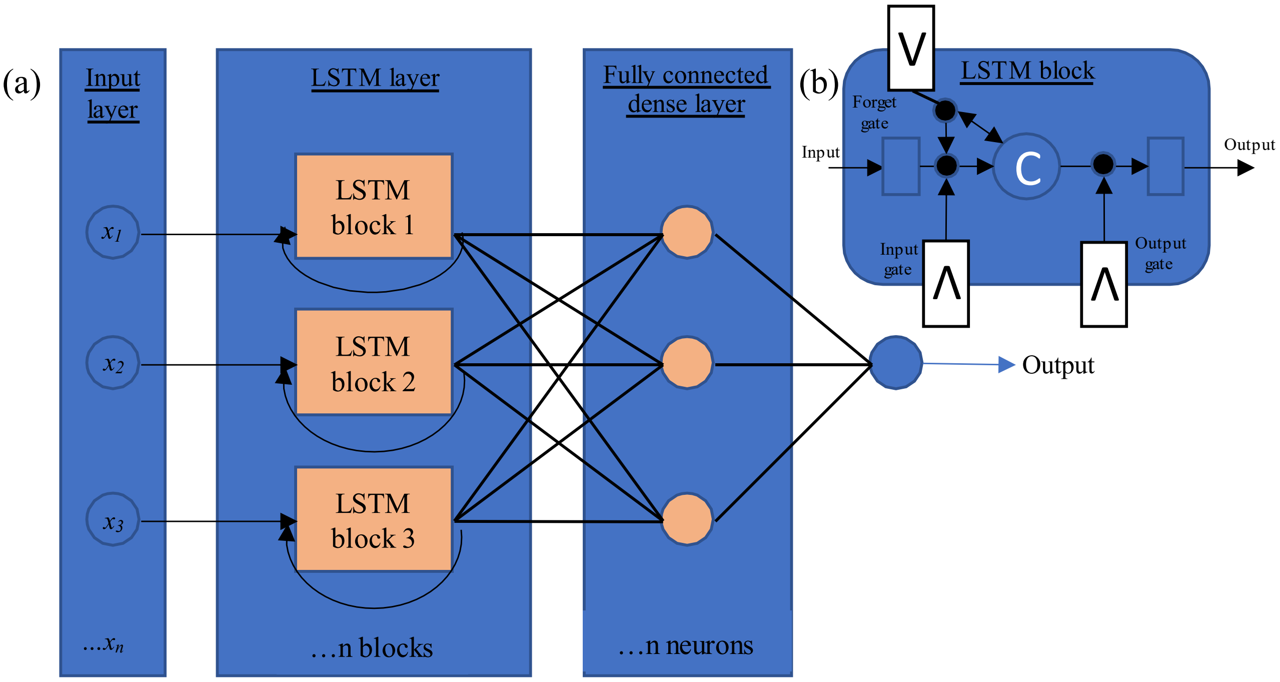

Figure 3. (**a**) simple schematic of an LSTM based architecture. Indicated is the input layer with input vectors $x_1 - x_n$. The hidden layer is in the middle, with the LSTM blocks. Followed by a single fully connected layer providing output; and (**b**) simple internal schematic of an LSTM block. Here the input, output, and forget gates control the flow of information into and out of the memory cell (C).

Figure 3 illustrates a schematic of a basic LSTM network and associated LSTM block. As with other recurrent neural networks, the architecture of LSTM consists of an input layer, hidden layers, and an output layer. The LSTM blocks typically recurrently connected memory units, replacing neurons in the hidden layer [51]. In addition, recurrent neural networks generally have feedback connections instead of conventional feedforward networks [48,50]. LSTMs can be trained in a supervised manner using training data, gradients loss function and back-propagation of the error at the output layer of the LSTM. Each weight of the LSTM network is optimised using a function of the derivative of the error [48]. This study relies on the Keras implementation of LSTM networks built on Tensorflow (https://keras.io/about/, (accessed on 11 July 2022)).

3.2. Data

This study chose six interrelated hydroclimatic, land-surface, and sub-surface parameters to derive a set of input features. These include groundwater storage anomaly, soil moisture, evapotranspiration, precipitation, runoff, and land surface temperature. Together, these parameters are considered major hydrological and hydrogeological components affecting the terrestrial water system in the case study area. They are thus chosen to predict a single output feature—groundwater level changes. Table 1 provides a summary of the parameters collected, their respective dataset sources and the final post-processed form before integration.

Table 1. Summary of the parameters data sources and their post-processing form.

Parameters	Source	Post-Processed Specifications		
		Temporal Resolution	**Spatial Resolution**	**Units**
Soil moisture	ERA5-Land hourly data from 1950 to present [1]	Monthly (average)	0.1° × 0.1°	centimetres
Precipitation	ERA5-Land hourly data from 1950 to present [1]	Monthly (cumulative)	0.1° × 0.1°	centimetres
Run-off	ERA5-Land hourly data from 1950 to present [1]	Monthly (cumulative)	0.1° × 0.1°	centimetres
Evapotranspiration	ERA5-Land hourly data from 1950 to present [1]	Monthly (cumulative)	0.1° × 0.1°	centimetres
Land surface temperature	ERA5-Land hourly data from 1950 to present [1]	Monthly (average)	0.1° × 0.1°	kelvin
GRACE ΔTWS	CSR GRACE/ GRACE-FO RL06 Mascon Solutions (version 02) [2]	Monthly	0.1° × 0.1°	centimetres, ΔGWS
GLDAS SM	GLDAS Noah Land Surface Model L4 monthly 0.25 × 0.25 degree V2.0 [3]	Monthly	0.25° × 0.25°	centimetres
GLDAS CW	GLDAS Noah Land Surface Model L4 monthly 0.25 × 0.25 degree V2.0 [3]	Monthly	0.25° × 0.25°	centimetres
GLDAS SWE	GLDAS Noah Land Surface Model L4 monthly 0.25 × 0.25 degree V2.0 [3]	Monthly	0.25° × 0.25°	centimetres
Groundwater levels	Malawi Ministry of Forestry and Natural Resources	Monthly	-	centimetre, ΔGWL

[1] https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview (accessed on 11 July 2022). [2] http://www2.csr.utexas.edu/grace/RL06_mascons.html (accessed on 11 July 2022). [3] https://disc.gsfc.nasa.gov/datasets?keywords=%20GLDAS&sort=spatialRes&page=1&source=Models%20Noah-LSM (accessed on 11 July 2022).

3.2.1. GRACE and Terrestrial Water Storage Anomalies

The Gravity Recovery and Climate Experiment (GRACE) mission consist of twin satellites working in tandem to measure changes in Earth's gravity field [52–54]. Measurements collected by the satellite instrumentation are put through various post-processing steps to convert satellite data into a number of data products, such as monthly changes in terrestrial water storage (ΔTWS) or monthly changes in ocean bottom pressure [55]. For example, removing the gravitational effects of the atmosphere, solid Earth and ocean tides, glacial isostatic rebound, and polar region ice mass, the signal change caused by variations in terrestrial water storage can be isolated [56]. For this study, a level 3 data product representing monthly changes in terrestrial water storage (ΔTWS) was retrieved from the Center for Space Research at the University of Texas [57]. This version (CSR GRACE/GRACE-FO RL06 Mascon solution version 02) covers a global spatial extent on a 0.25° × 0.25° grid, with ocean signals masked out. The coverage period extends from April 2002 to March 2020 (http://www2.csr.utexas.edu/grace/RL06_mascons.html, (accessed on 11 July 2022)).

3.2.2. GLDAS NOAH Derived Terrestrial Water Storage Data

The Global Land Data Assimilation System (GLDAS) is a state-of-the-art data assimilation and land surface modelling system designed by NASA to optimally simulate various land surface states at various temporal and spatial scales [58]. In this study, several GLDAS derived land surface variables were retrieved from NASA Goddard Earth Sciences Data and Information Services Center from April 2002 to March 2020 (https://disc.gsfc.nasa.gov/datasets?keywords=GLDAS, (accessed on 11 July 2022)). The version acquired for this study is the GLDAS Noah Land Surface Model L4 monthly 0.25 × 0.25-degree V2.0 (GLDAS_NOAH025_M 2.0) data product. It provides data as global monthly averages on a 0.25° × 0.25° grid. The variables retrieved include soil moisture (SM), canopy water storage (CW), and snow water equivalent thickness (SWE) fields.

3.2.3. ECMWF ERA5-Land Variables

The European Centre for Medium Weather Forecast's (ECMWF) ERA5 Land is the latest iteration of their reanalysis dataset developed through an advanced land surface modelling system. ERA5-Land provides data on several land surface states and single-level atmospheric variables, such as precipitation. In this study, several variables were retrieved, covering the extent of the study area, from the Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/home, (accessed on 11 July 2022)). The period of coverage extends from January 2000 to December 2019. These variables include soil moisture (four layers: 0–7 cm, 7–28 cm, 28–100 cm, and 100–289 cm), runoff (surface and shallow sub-surface), precipitation, evapotranspiration, and land surface temperature. Together these variables provide important contributions to the terrestrial energy and water budget for a number of aquifer settings, including the Shire Valley Alluvial Aquifer. The datasets are provided on an approximately 9 km × 9 km grid, with measurements at an hourly interval.

3.2.4. In-Situ Groundwater Level Measurements

As mentioned previously, the Shire Valley Alluvial Aquifer does not have an extensive groundwater level data record. There are only three long term monitoring boreholes in the study area (Figure 4). Furthermore, the time series for these boreholes only contain a limited number of observations, with the earliest record starting in 2009, and only a single borehole (Ngabu) has observations that extend until 2019. Furthermore, no data could be acquired for the Mozambique side of the aquifer. The available data were acquired courtesy of the Malawi Ministry of Forestry and Natural Resources.

3.3. Data Pre-Processing and Feature Engineering

As mentioned in Section 3, to train the machine learning algorithms, several input features (hydroclimatic variables) were chosen to model a single output or target feature (groundwater level anomalies). The following sections provide details of the pre-processing steps taken to transform the raw data into the covariate input and output features.

3.3.1. GRACE Groundwater Storage Anomaly

The version of GRACE data used in this study does not require additional post-processing, such as the application of de-stripping filters or optional signal gain factors. This allows direct application in various hydrosphere related studies [57]. However, ΔTWS data represent a combined value of the total change in the entire terrestrial water column (i.e., water stored as groundwater, soil moisture, canopy water storage, snow-water storage and surface water bodies). To extract the groundwater signal from GRACE data, the rest of the components of the terrestrial water budget must be removed from the total GRACE signal [59]. This can be done using the following terrestrial water budget formula [59]:

$$\Delta\text{GWS} = \Delta\text{TWS} - \Delta(\text{SM} + \text{SWE} + \text{CW}) \tag{1}$$

where ΔGWS is the groundwater storage anomaly, ΔTWS is the terrestrial water storage anomaly, and Δ(SM + SWE + CW) is the combined soil moisture (SM), snow-water equivalent (SWE) and canopy water storage anomaly (CW).

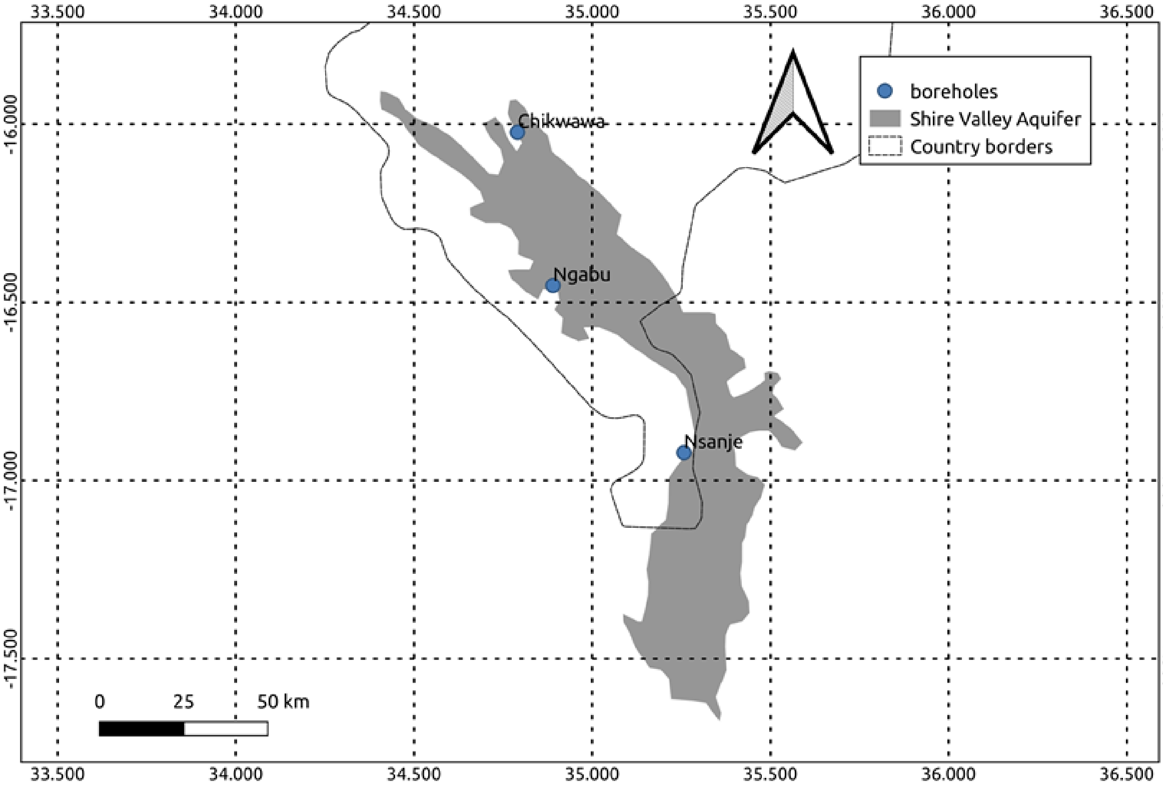

Figure 4. Groundwater level monitoring points within the Shire Valley Alluvial Aquifer.

Before executing Equation (1), a few pre-processing steps must be applied to the data. Firstly, gaps in the GRACE ΔTWS time series were filled by calculating average values of GRACE ΔTWS for the 12 calendar months and substituting these values at corresponding gaps in the time series. Secondly, the combined value for soil moisture, snow water equivalent thickness and canopy water storage anomalies Δ(SM + SWE + CW) must be calculated. To accomplish this task, datasets retrieved from the GLDAS data were used (Section 3.2.2). The units (kg/m^2) for SM, SWE and CW from GLDAS were converted to centimetres. After that, the individual layers were aggregated through summation into a combined land surface water budget component (SM + SWE + CW). To calculate Δ(SM + SWE + CW), the average monthly values of SM + SWE + CW between the period 2004–2009 were calculated. This period represents the same baseline period used to calculate GRACE ΔTWS. The average monthly SM + SWE + CW is then subtracted from each corresponding value in the time-series to derive Δ(SM + SWE + CW). This new value represents the deviation from the baseline period. It is important to note that surface water reservoirs and biomass can influence the GRACE signal. However, due to a lack of compatible datasets, they have not been included in the above formula [59].

The penultimate step then involved subtracting the Δ(SM + SWE + CW) from the GRACE ΔTWS to derive the GRACE ΔGWS. Finally, a bilinear spatial interpolation technique was used to transform the gridded ΔGWS data from 0.25° × 0.25° to a resolution of 0.1° × 0.1° [60] to match that of the ERA5-Land variables.

3.3.2. ERA5-Land Data Processing

The variables retrieved from ECMWF's ERA5-Land reanalysis dataset (Section 3.2.3) were subject to a number of pre-processing steps. All units were first converted to centimetres, except for land-surface temperature, which was maintained as units of kelvin. Before aggregating the data, the four soil moisture layers were combined into a single soil moisture layer by adding the individual layers together. Thereafter the values for all the variables were first aggregated into daily values, and then monthly values as follows: for precipitation, evapotranspiration, and runoff, the total accumulation was calculated, while for soil moisture, and land surface temperature, the mean value was calculated.

3.3.3. Calculating Groundwater Level Changes and Predictant Feature Generation

The time-series data were first aggregated into average monthly depth to groundwater level measurements for the three boreholes in the Shire Valley Alluvial Aquifer. After that, the monthly changes in depth to groundwater level were calculated by subtracting subsequent months from each other. This dataset represents the target feature used during model training and testing.

3.3.4. Data Integration and Predictor Feature Generation

The processed regional hydroclimatic data had to be extracted for each borehole to generate the set of input features. Using the coordinates of each borehole, the monthly data from the underlying pixel were extracted for the respective months in the time series. thereafter, for experiment I, 13 input features (specifically, the current month's hydroclimatic inputs, the previous month's hydroclimatic inputs and the previous month's groundwater level change) and one output feature (the current month's groundwater level change) were created by shifting the time series accordingly. Whereas for experiment II, 14 input features (specifically, the current month's hydroclimatic inputs, the previous month's hydroclimatic inputs, as well as the current and previous month's groundwater level change), and one output feature (the following month's groundwater level change) were created by shifting the time series accordingly. Table 2 summarises a full list of the input features for both experiments' time series. Pearson's product–moment correlation analysis was performed on the input and output features, for each borehole and experiment, to understand the potential relationships between covariates. Finally, all the values in the dataset were normalised to between 0 and 1.

Table 2. Summary of predictor and predictant features used for model training and testing.

Experiment	Feature Labels	Description	Feature Category
I/II	precip (t)	Total precipitation for the current month	Inputs
	precip (t − 1)	Total precipitation for the previous month	
	evap (t)	Total evapotranspiration for the current month	
	evap (t − 1)	Total evapotranspiration for the previous month	
	ro (t)	Total runoff for the current month	
	ro (t − 1)	Total runoff for the previous month	
	sm (t)	Average soil moisture for the current month	
	sm (t − 1)	Average soil moisture for the previous month	
	lst (t)	Average land-surface temperature for the current month	
	lst (t − 1)	Average land-surface temperature for the previous month	
	ggwsa (t)	Current months GRACE GWS anomaly	
	ggwsa (t − 1)	Previous months GRACE GWS anomaly	
	delta (t − 1)	Groundwater-level change for the previous month	
	delta (t)	Groundwater-level change for the current month	Input (II)/output (I)
II	delta (t + 1)	Groundwater-level change for the following month	Output

3.4. Tuning, Training and Testing

The time series was split along the time axis into training and testing sets for each borehole. This was done to maintain the 'series' sequential structure. The size and location of the training and test sets were based on each borehole's unique time-series distribution—for example, gaps in the data had to be considered when selecting the training and test sets. For specifications on each experiment's train-test split, please refer to Section 4.1. Due to the limited sample size, no validation set was employed during training.

3.4.1. GBDT Model

Hyperparameter tuning was undertaken using a 10-fold cross-validation technique applied through Scikit-learns' gridsearchCV function, using root mean square error (RMSE) as a scoring metric. Only the most basic hyperparameters were selected for tunings, such as maximum depth, minimum child samples, number of estimators, and number of leaves. Tuning was carried out using only the training set. Table 3 summarises the configurations and hyperparameter settings tested during the gridsearch. Scikit-learns' gridsearchCV function tests each of the possible permutations of hyperparameter settings across those provided in Table 3 and returns the hyperparameter settings for the estimator with the lowest overall RMSE. The maximum values for each of the parameters were kept small, due to the limited number of available samples, and to avoid growing large trees, that may cause overfitting. Regularisation was not considered, and the learning rate was kept constant at 0.1 across all test cases. Scoring the estimators in the gridsearch and selecting the best parameter setting was done using the RMSE metric. The lower the value (closest to zero) for this metric, the better the predictive model. The final model was then trained and tested on the training and test dataset, respectively, with the performance reported in RMSE and mean absolute error (MAE), coefficient of determination (R^2) and Nash–Sutcliffe model efficiency coefficient (NSE). RMSE was again used as the loss function to train the final GBDT model.

Table 3. Configurations and hyperparameter settings used in the gridsearchCV.

Hyperparameter	Settings
Max depth	1, 2, 3, 4, 5
Min. child samples	1, 2, 4, 8, 16
No. of estimators	10, 20, 40, 80, 160
No. of leaves	1, 2, 4, 8, 16, 32

3.4.2. LSTM Model

Hyperparameter tuning was performed only on the most basic model parameters, including the number of epochs, batch size and the number of neurons. Like with the GBDT, no regularisation or drop-out was used to slow down the learning rates. Hyperparameter tuning for the LSTM models was done through empirical testing of various combinations of training parameter settings. Table 4 illustrates the combination of hyperparameter settings used during the tuning phase. Unlike with the GBDT model, not all permutations of hyperparameter settings were tested exhaustively due to the manner in which hyperparameter tuning was carried out for the LSTM model. In this case, the tuning process involved tuning and fixing the best setting for each hyperparameter successively, starting with the number of epochs, followed by the batch size, and lastly, the number of neurons. For example, once the optimal number of epochs was determined in the first stage, this value was fixed for the subsequent stages (i.e., batch size and the number of neurons). In Table 4, the values of the epochs reflect the optimal range of settings, based on prior testing (not shown here). The batch sizes were chosen, so as to be devisable by the sample sizes, in-order to maintain the statefulness of the LSTM model. Finally, the number of neurons were chosen to avoid complex hidden layers, that might not be ideal for small sample sizes.

Table 4. Description of the hyperparameter setting combinations tested in the LSTM model.

	Experiment I		Experiment II	
Hyperparameter	**Ngabu**	**Nsanje**	**Ngabu**	**Nsanje**
No. of epochs	100, 200, 300, 400, 600, 800, 1000	200, 400, 600, 800, 1000, 1200, 1600, 2000	200, 300, 400, 600, 800, 1000, 1200, 1400	100, 200, 400, 600, 800, 1000, 1200
Batch size	1, 2, 4, 8	1, 2, 3, 6	1, 2, 4, 8	1, 2, 4, 8
No. of neurons	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5	1, 2, 3, 4, 5

Due to the stochastic nature of LSTM algorithms, each combination of parameter settings was subject to 30 repeated iterations of model training and testing. The average model scores across the 30 iterations were returned and used to assess performance. Specifically, performance was assessed using the average and variance of the RMSE scores across the 30 iterations. Selecting the best combination of parameter settings is a trade-off between the lowest average and smallest variation in error scores. For every LSTM model, RMSE was used as the gradient loss function, with network weights updated using the adam optimiser. In addition, every LSTM network was built with a single dense layer comprising 1 unit after the LSTM layer. The overall best combination of parameter settings was then used to train and test the final model across 30 iterations, and the performance was reported in RMSE, MAE and R^2 and NSE.

4. Results

4.1. Groundwater Levels in the Shire Valley Alluvial Aquifer

Figure 5 illustrates the hydrographs for the three monitoring boreholes in the Shire Valley Alluvial Aquifer across the various pre-processing stages (i.e., aggregation and anomaly calculations). On the monthly scale, the groundwater level record for the Chikwawa borehole starts in 2010 and extends until 2013. However, this borehole was dropped from the experiments due to numerous time series gaps coupled with an already limited sample size (35 data samples). Ngabu, on the other hand, contains the greatest number of groundwater level records (n = 96) on the monthly scale. Here the record extends from 2009 to the end of 2019, albeit with a large gap from the years 2014 through to the end of 2016. Lastly, the Nsanje borehole contains 45 records on a monthly scale. The record for this borehole extends from 2010 to the end of 2013, with only a small gap at the end of the time series. Note that the sample sizes for the groundwater level anomalies are generally smaller than the monthly static groundwater level records, as numerous observations have no neighbouring observations needed to calculate the anomalies. These observations were thus dropped before creating the training and test datasets.

Table 5 shows the train and test set to split for each borehole and experiment. The training and test sets were identical for both GBDT and LSTM models. Additionally, the number of samples in each set was also chosen such that they were wholly divisible by the batch sizes. This was done to ensure the LSTM models remained stateful. For the Ngabu borehole, the training and test sets were defined, taking into account the gap in the time series. Specifically, the training set was populated with samples before the gap, while the test set only contained samples after the gap. As for the Nsanje boreholes, the time series was intact, albeit only with a limited number of observations. Here the time series was split conventionally along the time axis.

Table 5. Sample sizes for the train and test set to split for each experiment and borehole.

	Experiment I		Experiment II	
	Train	**Test**	**Train**	**Test**
Ngabu	56	32	56	32
Nsanje	30	12	32	8

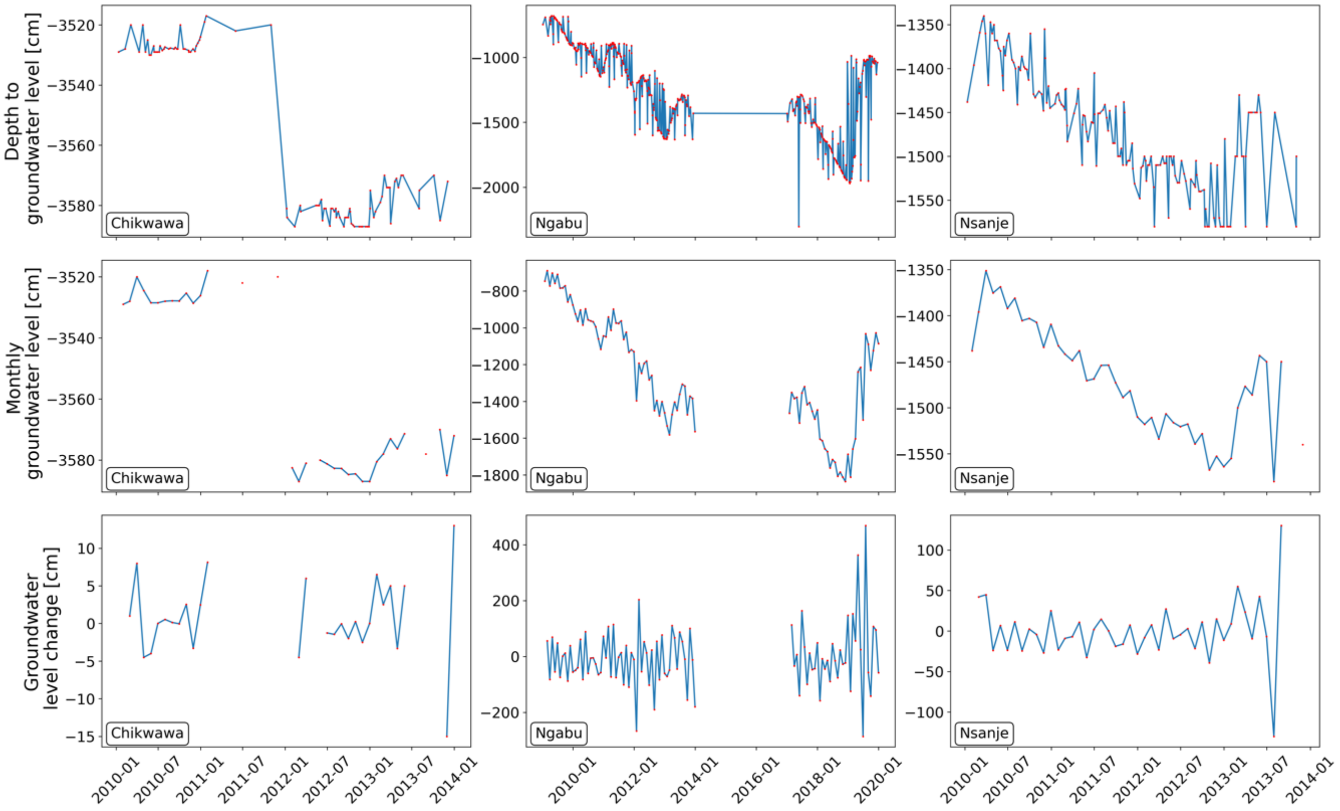

Figure 5. Hydrographs of the three monitoring boreholes in the Shire Valley Alluvial Aquifer. The first row illustrates the raw data, the second row shows the mean monthly groundwater levels, and the bottom row shows the monthly groundwater level changes.

4.2. Input and Output Features

Figure 6 illustrates the time series of the six hydroclimatic variables and the groundwater level anomalies. The input features have been shifted to align the previous time steps with the current or future groundwater level anomalies. The hydroclimatic variables demonstrate a typical seasonal cyclicity. This is also true for GRACE derived GWSA, which are closely related to groundwater level. The trends appear relatively similar between the two boreholes, albeit with a difference in the range of values. For example, precipitation and runoff are greater in the Ngabu borehole region than in the Nsanje borehole region. Finally, the groundwater level anomalies vary from month to month and have a minimal correlation to the hydroclimatic variables. Please note that due to the ECWMF Integrated Forecasting System's convention that upward fluxes are negative, the evapotranspiration values expressed in this study have negative values.

4.3. Pearson's Correlations Analysis

Here we inspect the correlation matrix for meaningful linear relationships between groundwater level anomalies (target) and the input features (Figure 7). Generally speaking, Pearson's correlation coefficients between groundwater level anomalies and the input variables range between -0.4 and 0.4. The variables with the greatest correlations are soil moisture, the previous month's groundwater level anomaly, precipitation and runoff inputs. Interestingly, the target variable has an inverse relationship with the previous month's groundwater level anomaly. The remainder of the input features, including GRACE groundwater storage, have insignificant correlations with the target feature.

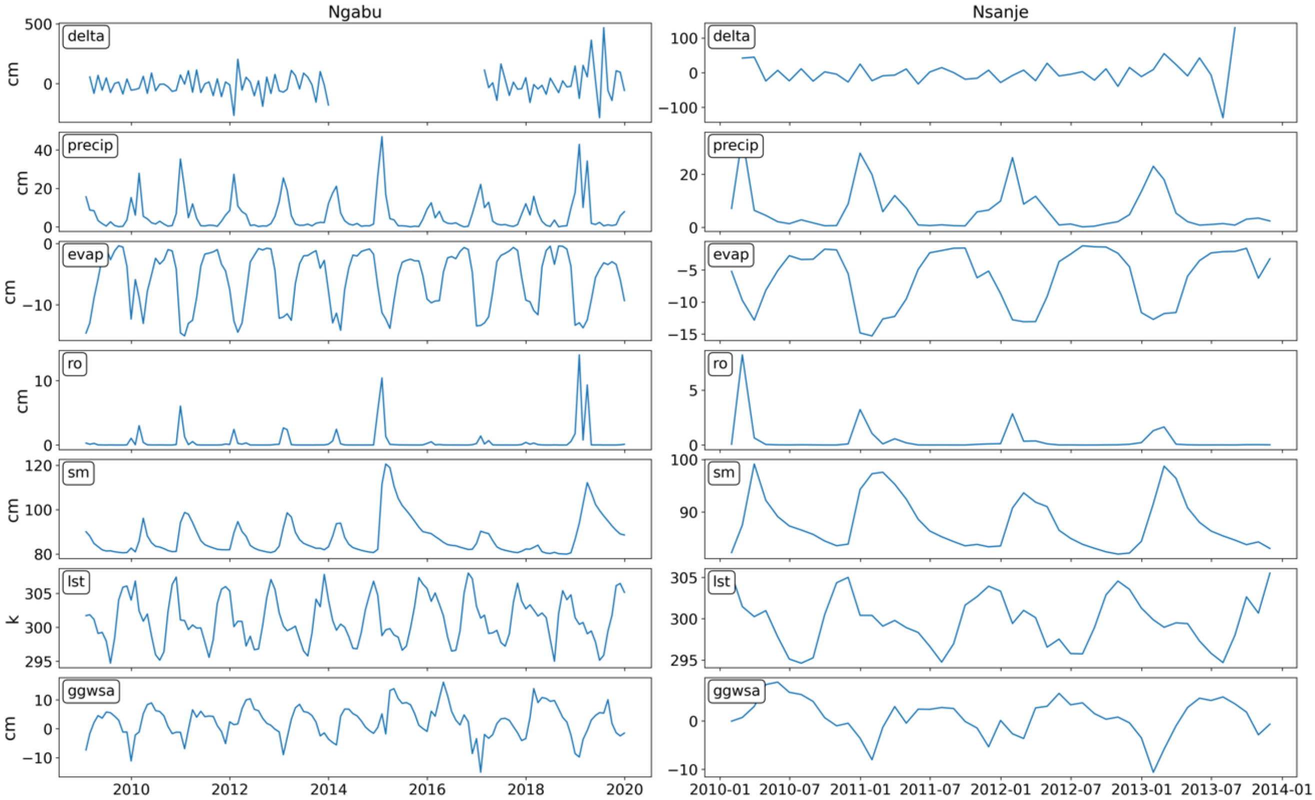

Figure 6. Time series of the various hydroclimatic inputs vs. groundwater level changes for each borehole. Delta = Monthly groundwater level change, precip = total monthly precipitation, evap = total monthly evapotranspiration, ro = total monthly runoff, sm = average monthly soil moisture, lst = average monthly land surface temperature, ggwsa = monthly GRACE groundwater storage change.

4.4. Hyperparameter Tuning

4.4.1. GBDT Models

Figure 8 illustrates the cross-validation curves in the hyperparameter space for the various hyperparameter combinations. The mean value curves indicate the average for all estimators using a particular hyperparameter setting, while the shaded area indicates 1 standard deviation on either side of the mean. Values closer to zero indicate better cross-validation scores on average. The mean value curves for the cross-validation score generally indicate minimal differences across the hyperparameter space. Significant overfitting occurs after a maximum depth of 1 for all experimental cases. For minimum child samples per leaf, the training and cross-validation curves appear to converge the greater the number of child samples per leaf, albeit sometimes with lower overall RMSE scores. For the number of estimators, significant overfitting occurs with an increase in boosting rounds. However, the cross-validation score tends to remain the same. This is true in all cases except for the Ngabu borehole in experiment I. The same is true for a number of leaves, where significant overfitting occurs after two leaves. In addition, there is generally a high standard deviation observed for training for each parameter setting across all the experimental cases. At the same time, lower standard deviations are observed for the cross-validation scores.

Scikit-learns' gridsearchCV functionality return the hyperparameter setting for the estimator with the best RMSE score. Table 6 indicates the parameters used for the final models. Here the maximum depth and number of leaves across all the experimental cases, appear representative of the hyperparameter space (Figure 8). However, both minimum child samples and the number of estimators did not strictly follow the trends observed in the hyperparameter space. While not shown here, a number of estimators produced identical scores with different minimum child samples and the number of estimators in use.

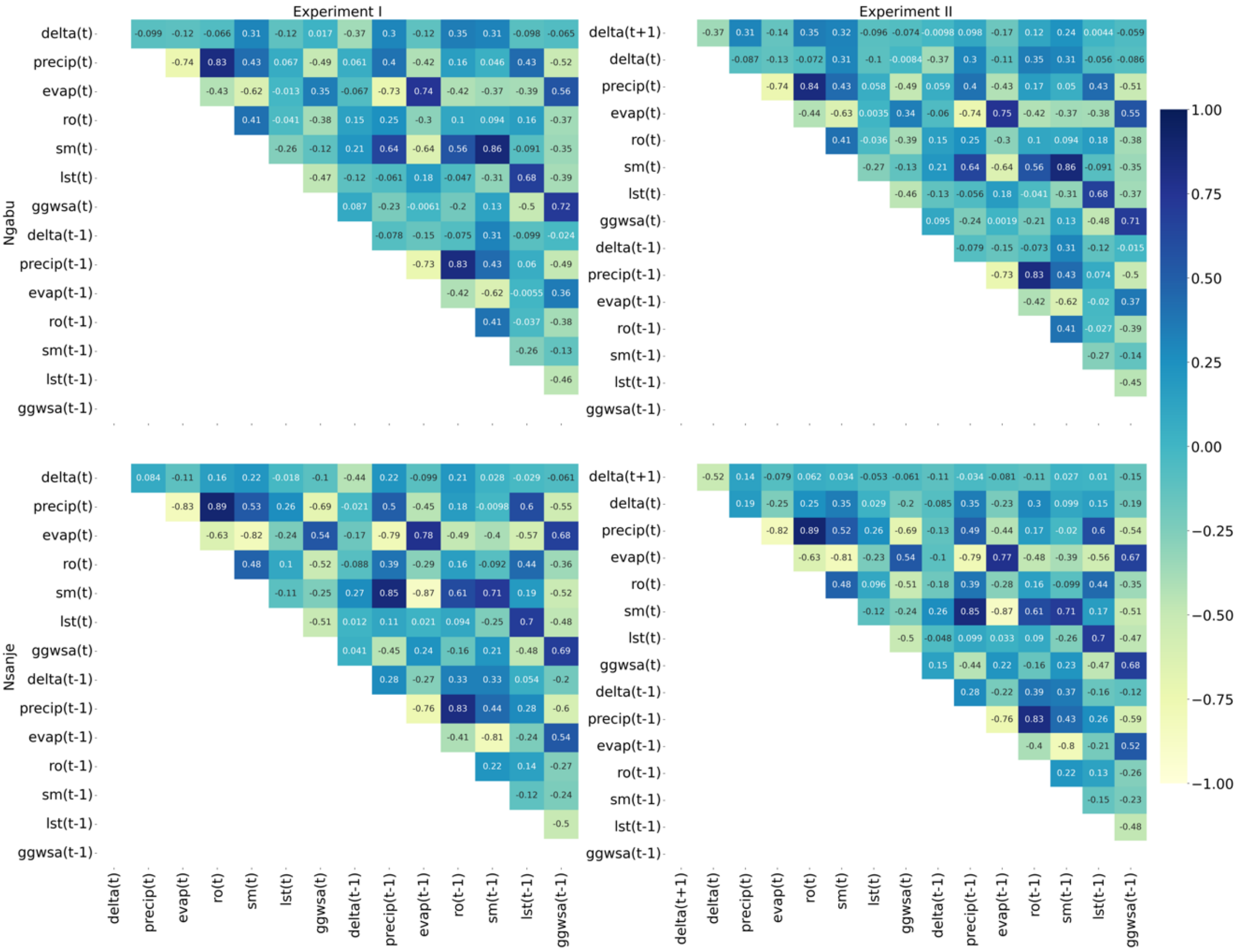

Figure 7. Pearson's product–moment cross-correlation plots for each of the input features and the target feature. The correlation coefficients are displayed on the plot. Please refer to Table 2 for a description of the labels.

Table 6. Optimal hyperparameter settings determined through the gridsearch 10-fold cross-validation.

	Experiment I		Experiment II	
Hyperparameter	**Ngabu**	**Nsanje**	**Ngabu**	**Nsanje**
Maximum depth	1	1	1	1
Minimum child samples	8	8	16	8
Number of estimators	40	80	80	160
Number of leaves	2	2	2	2
Error score (RMSE)	−0.0872	−0.0652	−0.0850	−0.0480

4.4.2. LSTM Models

Figure 9 illustrates the distributions of error scores for various hyperparameter settings during the tuning of the LSTM models. The distribution is calculated using the individual error scores attained during the 30 iterations per combination of hyperparameter settings. The preference in choosing the correct settings for the final model is a trade-off between average error scores and the distribution in error scores. The spread in the distribution indicates the uncertainty in model performance, and the average error score indicates the most likely prediction score. Low standard deviations illustrate better consistency in model performance, while lower average error scores indicate better model performance.

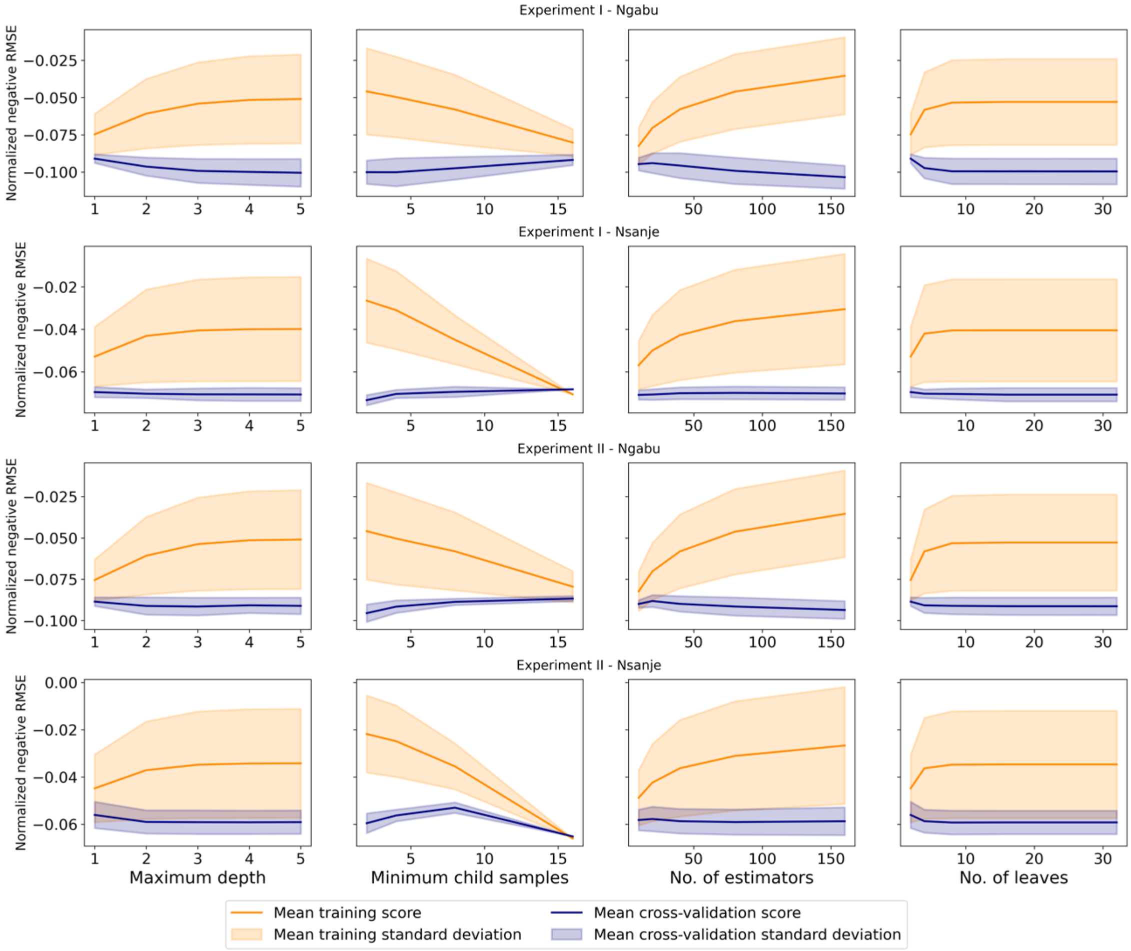

Figure 8. Plots of the cross-validation analysis results for each of the hyperparameters in the GBDT models across the various experiments. Note that the dimensions for error scores are returned as normalised negative RMSE values, which is simply the negated RMSE (i.e., unitless).

Compared to the GBDT model, the hyperparameter space for the LSTM models is complicated mainly by the stochastic nature of the LSTM model. This is evident by the inconsistency in trends within the hyperparameter space, especially for the same borehole, where the data are largely the same. In general, an increase in the number of epochs comes with an increase in uncertainty, albeit sometimes with a better average score. For the number of neurons, it is apparent that increasing the number of neurons in the hidden layer comes with an increase in uncertainty and a reduction in model performance. However, there appear to be no clear trends across the experimental cases for batch size, and it is mainly random.

Table 7 shows the final parameter settings for each of the experiments. These are based on the best performing estimators across the hyperparameter settings. Here, the epoch size chosen was reflective of the hyperparameter space. There appears to be better performance of models within the median range of epoch sizes for all cases, except for the Nsanje borehole in experiment II. Here, the lowest epoch setting appeared to have the greatest skill. Smaller batch sizes demonstrate better skill, except for the Ngabu borehole in experiment II, where the optimal setting was 8. The batch size chosen falls within the mini-batch learning regime in each case. The optimal number of neurons was set at 1 for all cases, except for the Ngabu borehole in experiment II. Here 4 neurons represented the optimal setting.

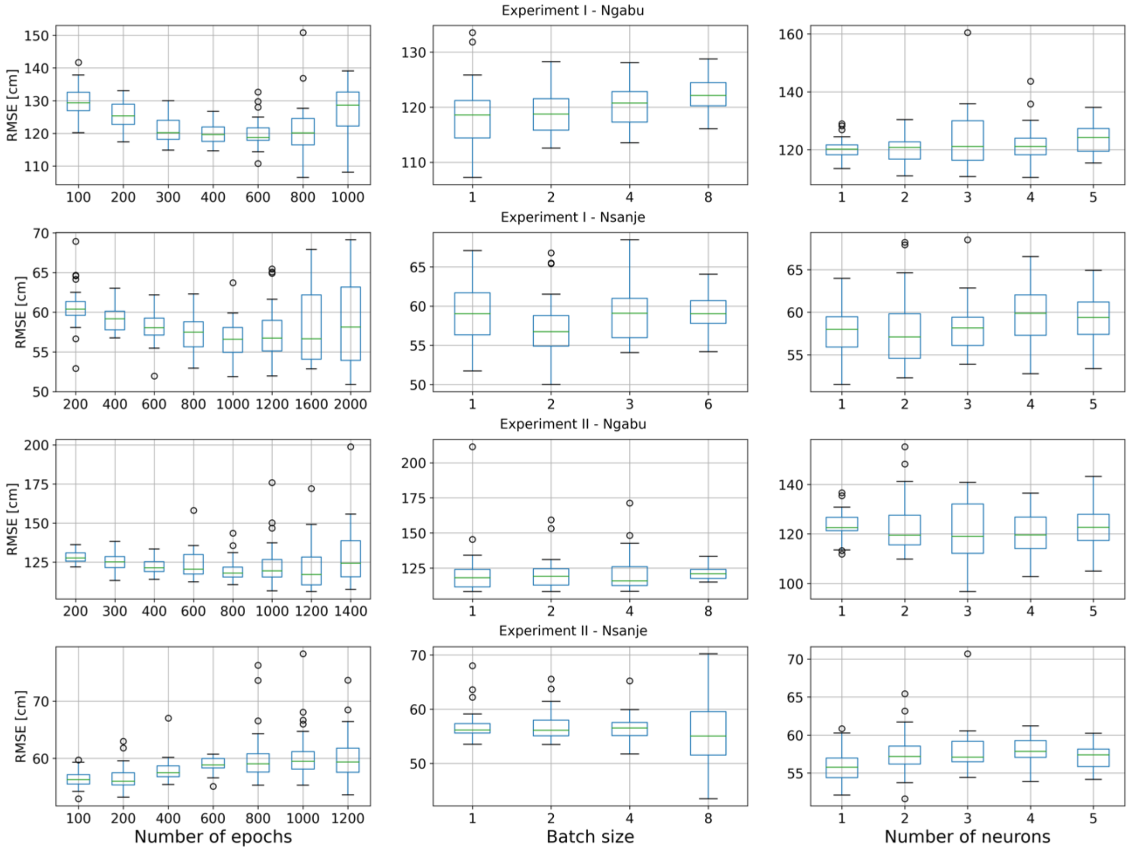

Figure 9. Box & whisker plots for each of the hyperparameters in the LSTM models across the various experiments.

Table 7. Optimal hyperparameter settings for each of the experiments.

	Experiment I		**Experiment II**	
Hyperparameter	**Ngabu**	**Nsanje**	**Ngabu**	**Nsanje**
Number of epochs	400	1000	800	100
Batch size	2	2	8	4
Number of neurons	1	1	4	1
Error score (RMSE)	120.11	57.58	119.98	55.80

4.5. Comparison of Final Model Predictions

Table 8 describes the test scores for each machine learning model and each of the experiments. By assessing the performance across the two boreholes, it can be seen that the prediction results of the Ngabu borehole were better (higher R^2 and NSE score) than those of the Nsanje borehole. This is true across both experiments and between both models. Results for experiment I were generally better than those of experiment II. The exception is LSTM models for Ngabu boreholes, indicating mainly equal or better prediction skills for experiment II than experiment I. When the results are compared across models, the LSTM model outperforms the GBDT model across all experiments and metrics, except for the Nsanje borehole in experiment II, were the NSE is better for GBDT.

Figure 10 illustrates the hydrographs of the predicted vs. observed groundwater level changes for each model. The GBDT model outperforms the LSTM for the Nsanje borehole over the training set, and the LSTM does slightly better over the test set. In fact, for the Nsanje borehole in experiment II, the LSTM remains largely conservative and fails to learn the groundwater level changes. When it comes to the Ngabu borehole, which has almost double the amount of data, the LSTM model outperforms the GBDT model in both the training and test set across both experiments. In conclusion, LSTM is generally better at predicting extreme groundwater level changes.

Table 8. Comparison errors scores between the GBDT and LSTM model on the test set for all the experiments.

	GBDT								**LSTM**							
	Ngabu				**Nsanje**				**Ngabu**				**Nsanje**			
	MAE	**RMSE**	**R^2**	**NSE**	**MAE**	**RMSE**	**R^2**	**NSE**	**MAE**	**RMSE**	**R^2**	**NSE**	**MAE**	**RMSE**	**R^2**	**NSE**
Experiment I	84.06	127.87	0.18	0.32	39.56	59.47	−0.04	0.12	82.20	118.83	0.30	0.40	42.46	57.44	0.03	0.18
Experiment II	92.10	134.72	0.14	0.27	46.55	58.35	−0.21	0.15	86.38	118.74	0.30	0.44	39.24	56.81	−0.15	−0.09

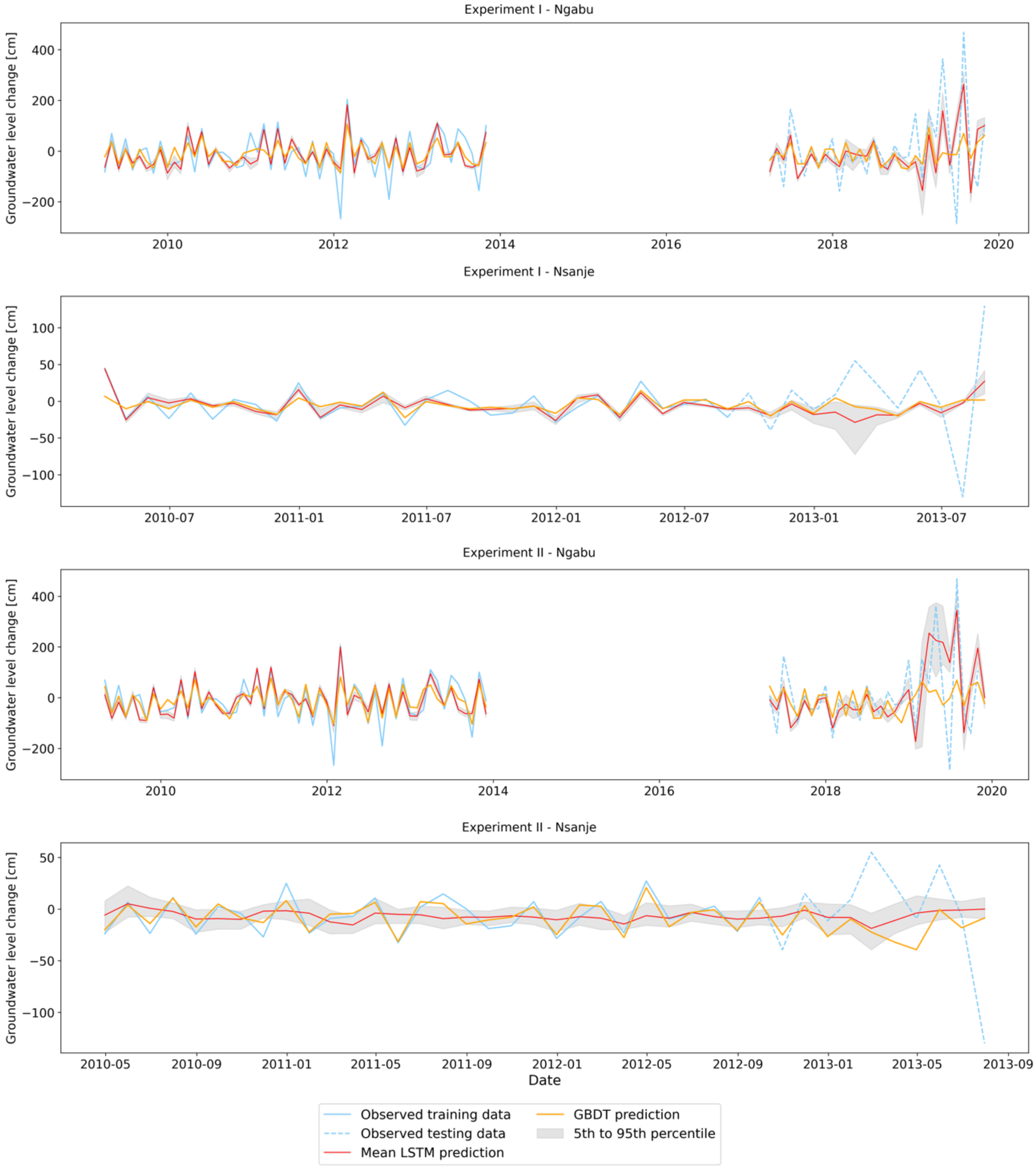

Figure 10. Hydrographs of the predicted vs. observed groundwater level changes for each experiment.

5. Discussion

The groundwater level measurement recorded for the three long-term monitoring boreholes contained only limited observations, with numerous gaps in some cases. One borehole, Chikwawa, had to be dropped from the experimentation due to limited records and numerous gaps, which would not be conducive to model training. These gaps would make it challenging to define a coherent successive sample for the training and testing datasets. Like the Chikwawa boreholes, the Ngabu also contained a gap in the time series, however, in this case, the single gap was catered for by assigning those records before the gap into the training set and those records after the gap into the test set. While this is not ideal, the series' sequential nature is primarily maintained, albeit with a gap. The Nsanje borehole had no gaps in the data but, on the other hand, had a small sample size.

It is also important to note that for both the Ngabu and Nsanje boreholes, the distributions for the test set appear to be skewed away from the training set distribution due to extreme values. Training and testing on different distributions, can have a significant impact on the model's performance.

Regarding the input features and their importance in the predictive skills, we see that the input features generally correlate poorly with the target feature. The target feature appears to be influenced the most by the previous month's groundwater level change. Other variables such as precipitation, soil moisture and runoff provide some explanation for the trend in the target feature. After all, these three features are interrelated and point to some rainfall-induced recharge. Interestingly, the GRACE groundwater storage anomaly demonstrates almost no correlation with monthly groundwater changes, which is surprising considering the close relationship between groundwater storage and groundwater levels. It is important to note that the lack of correlation between the input features and the target feature may result from the spatial and temporal resolutions of the data. Generally speaking, fluctuations of groundwater levels in a borehole are driven mainly by local processes, which may not be able to be resolved by the coarseness of the input data. In this regard, additional features may be needed to provide predictability of the target feature further. Especially with regards to extreme fluctuations, primarily a result of groundwater abstraction.

Finally, the tuning, training, and testing of the GBDT and LSTM models, allow a comparison of various machine learning models in data-scarce environments. Starting with the tuning process, the GBDT requires far less effort and time to tune compared to the LSTM model. The LSTM model is significantly slower to tune and train, which is typical among deep learning models. Furthermore, selecting the optimal hyperparameter setting was often challenging due to the stochastic nature of the LSTM and how the batch size is selected (online vs. full batch) introduce randomness into the outcomes. Nonetheless, if correctly tuned and trained, the LSTM provides better skill in predicting groundwater level changes. In addition, the LSTM allows for probabilistic outputs, where in some cases, the individual predictions are closer to the observed value than the mean of the outputs.

However, there may be a point (in terms of sample size) below which it becomes impractical to develop an LSTM model for our particular dataset. This is certainly the case when reviewing the results for the Nsanje borehole. Here, the increase in performance is negligible compared to the cost of tuning and training the LSTM model. It may be better to rely on GBDT models or other classical machine learning algorithms. As soon as the size of the training and test sets are increased (in this case, twice as many as that of Nsanje borehole), the LSTM model outperforms the GBDT model. In essence, more data are needed to develop better-generalised models.

Compared to results from previous studies, there appears to be little consensus with the results achieved in this study. For example, Wunsch [26] concluded that a shallow neural network (NARX) was more efficient and better at predicting groundwater levels for a long time series. However, when previous time steps are included in the input vector, the deep learning models outperform the NARX algorithm. This scenario resembles closer the setup of the experiments in this study and appears to confirm our findings.

The results are underwhelming when it comes to the actual performance of the models in this study compared to previous studies. For instance, Malakar, Wunch, Solgi, Afan and Guo [25,26,39,61,62], all report far superior performance scores across several metrics for deep learning models including LSTMs. Even the GBDT model in this study demonstrated much lower performance scores than previous studies using classical and ensemble machine learning approaches. For example, Seyoum, Kombo, Milewski and Malakar [18,33,38,63] all demonstrate superior model performance on individual time series predictions using classical machine learning algorithms, including GBDT. However, significantly more data were available and used to train and test the machine learning algorithm in previous studies. This may suggest that more data, especially ground-based, are needed to develop robust and generalised models and, in essence, reaffirm the conclusions of [12].

6. Conclusions

The goal of this study was to compare the usefulness of these two machine learning algorithms, GBDT and LSTM, as a helpful tool to model groundwater level changes in data-scarce aquifer (Shire Valley Alluvial Aquifer), which is typical of SADC settings. The results illustrate that the LSTM model performed better than the GBDT model in most experiments, especially in response to extreme groundwater level changes. However, upon closer inspection, the GBDT model appeared to provide only slightly worse skill for the Nsanje borehole in experiment I and much better skill for the Nsanje borehole in experiment II, though in this case, the Nsanje borehole had a smaller sample size. These results point out that, in most cases, the LSTM may be the most appropriate model to use when more data are available, but that there may be a case for the GBDT model when only small sample sizes are available. For such small datasets, the cost of tuning, training, and testing and LSTM for minimal gains may not be justified. In conclusion, while there may be a case for machine learning algorithms in data-scarce aquifers of SADC, more data are needed to improve the reliability of the models.

This study does not constitute an exhaustive assessment of the problem area. Hence future research should explore experimentation with various other machine learning algorithms, such as transfer learning and federated learning approaches. Further work could also explore different architectures such as additional hidden layers, different nodal density per layer, various loss functions, or the use of drop-out and regularisation, which should improve learning. Furthermore, to better capture the trends in groundwater level changes, additional input features could be needed, such as groundwater abstractions or the influence of surface water groundwater interaction.

Author Contributions: Conceptualization, Z.G., K.P. and A.B.; methodology, Z.G., A.B. and O.A.; software, Z.G.; validation, K.P., N.J., A.B. and O.A.; formal analysis, Z.G.; investigation, Z.G., N.J., A.B. and O.A.; resources, Z.G.; data curation, G.W.; writing—original draft preparation, Z.G.; writing—review and editing, K.P., N.J., A.B., O.A., T.K. and G.W.; visualisation, Z.G.; supervision, K.P., T.K., A.B. and N.J.; project administration, K.P., T.K. and A.B.; funding acquisition, K.P., A.B. and T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, a multi-agency coalition with the Department of Science and Innovation South Africa, the USAID, the Southern African Development Community-Groundwater Management Institute (SADC-GMI), the South African Water Research Commission (WRC), the Department of Science and Innovation South Africa, the Sustainable Water Partnership, IBM Research Africa, and the United States Geological Survey (USGS). Further financial assistance towards APC was provided by IBM Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the Groundwater Division of the Water Resources Department of the Ministry of Water and Sanitation, Malawi.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Tsai, C.-W.; Lai, C.-F.; Chao, H.-C.; Vasilakos, A.V. Big Data Analytics: A Survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
2. García, S.; Ramírez-Gallego, S.; Luengo, J.; Benítez, J.M.; Herrera, F. Big Data Preprocessing: Methods and Prospects. *Big Data Anal.* **2016**, *1*, 9. [CrossRef]
3. Raghupathi, W.; Raghupathi, V. Big Data Analytics in Healthcare: Promise and Potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef]
4. Roy, A.K. Advances and Scope in Big Data Analytics in Healthcare. *Curr. Trends Biomed. Eng. Biosci.* **2017**, *9*, 55758. [CrossRef]
5. Zhang, Y.; Zhao, Y. Astronomy in the Big Data Era. *Data Sci. J.* **2015**, *14*, 11. [CrossRef]
6. Guo, H. Big Data Drives the Development of Earth Science. *Big Earth Data* **2017**, *1*, 4–20. [CrossRef]
7. Mohammadpoor, M.; Torabi, F. Big Data Analytics in Oil and Gas Industry: An Emerging Trend. *Petroleum* **2018**, *6*, 321–328. [CrossRef]
8. Sudmanns, M.; Lang, S.; Tiede, D. Big Earth Data: From Data to Information. *GI_Forum* **2018**, *1*, 184–193. [CrossRef]
9. Chen, H.; Chiang, R.H.; Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.* **2012**, *36*, 1165. [CrossRef]
10. Adamala, S. An Overview of Big Data Applications in Water Resources Engineering. *Mach. Learn. Res.* **2017**, *2*, 10–18. [CrossRef]
11. Chen, Y.; Han, D. Big Data and Hydroinformatics. *J. Hydroinform.* **2016**, *18*, 599–614. [CrossRef]
12. Ahmadi, A.; Olyaei, M.; Heydari, Z.; Emami, M.; Zeynolabedin, A.; Ghomlaghi, A.; Daccache, A.; Fogg, G.E.; Sadegh, M. Groundwater Level Modeling with Machine Learning: A Systematic Review and Meta-Analysis. *Water* **2022**, *14*, 949. [CrossRef]
13. Tao, H.; Hameed, M.M.; Marhoon, H.A.; Zounemat-Kermani, M.; Heddam, S.; Kim, S.; Sulaiman, S.O.; Tan, M.L.; Sa'adi, Z.; Mehr, A.D.; et al. Groundwater Level Prediction Using Machine Learning Models: A Comprehensive Review. *Neurocomputing* **2022**, *489*, 271–308. [CrossRef]
14. Gaffoor, Z.; Gritzman, A.; Pietersen, K.; Jovanovic, N.; Bagula, A.; Kanyerere, T. An Autoregressive Machine Learning Approach to Forecast High-Resolution Groundwater-Level Anomalies in the Ramotswa/North West/Gauteng Dolomite Aquifers of Southern Africa. *Hydrogeol. J.* **2022**, *30*, 575–600. [CrossRef]
15. Johnny, J.C.; Sashikkumar, M.C.; Sivadevi, K.; Kirubakaran, M. Prediction of Groundwater Level Dynamics Using Artificial Neural Network. In Proceedings of the 2015 IEEE 7th National Conference on Computing, Communication and Information Systems, Coimbatore, India, 13–14 February 2015; The Shaheed Bhagat Singh State Technical Campus: Ferozepur, India, 2015; p. 7.
16. Kenda, K.; Čerin, M.; Bogataj, M.; Senožetnik, M.; Klemen, K.; Pergar, P.; Laspidou, C.; Mladenić, D. Groundwater Modeling with Machine Learning Techniques: Ljubljana Polje Aquifer. *Proceedings* **2018**, *2*, 697. [CrossRef]
17. Nayak, P.C.; Rao, Y.R.S.; Sudheer, K.P. Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. *Water Resour. Manag.* **2006**, *20*, 77–90. [CrossRef]
18. Seyoum, W.; Kwon, D.; Milewski, A. Downscaling GRACE TWSA Data into High-Resolution Groundwater Level Anomaly Using Machine Learning-Based Models in a Glacial Aquifer System. *Remote Sens.* **2019**, *11*, 824. [CrossRef]
19. Alahmadi, F.S. Groundwater Quality Categorization by Unsupervised Machine Learning in Madinah, Western Kingdom of Saudi Arabia. In Proceedings of the International Geoinformatics Conference 2019 (IGC2019), Nantes, France, 27–31 August 2019; International Society for Photogammetry and Remote Sensing: Riyadh, Saudi Arabia, 2019; p. 10.
20. Ransom, K.M.; Nolan, B.T.; Traum, J.A.; Faunt, C.C.; Bell, A.M.; Gronberg, J.A.M.; Wheeler, D.C.; Rosecrans, C.Z.; Jurgens, B.; Schwarz, G.E.; et al. A Hybrid Machine Learning Model to Predict and Visualize Nitrate Concentration throughout the Central Valley Aquifer, California, USA. *Sci. Total Environ.* **2017**, *601–602*, 1160–1172. [CrossRef]
21. Lee, S.; Hyun, Y.; Lee, M.-J. Groundwater Potential Mapping Using Data Mining Models of Big Data Analysis in Goyang-Si, South Korea. *Sustainability* **2019**, *11*, 1678. [CrossRef]
22. Hussein, E.A.; Thron, C.; Ghaziasgar, M.; Bagula, A.; Vaccari, M. Groundwater Prediction Using Machine-Learning Tools. *Algorithms* **2020**, *13*, 300. [CrossRef]
23. Raheja, H.; Goel, A.; Pal, M. Prediction of Groundwater Quality Indices Using Machine Learning Algorithms. *Water Pract. Technol.* **2022**, *17*, 336–351. [CrossRef]
24. Huang, X.; Gao, L.; Crosbie, R.S.; Zhang, N.; Fu, G.; Doble, R. Groundwater Recharge Prediction Using Linear Regression, Multi-Layer Perception Network, and Deep Learning. *Water* **2019**, *11*, 1879. [CrossRef]
25. Malakar, P.; Mukherjee, A.; Bhanja, S.N.; Sarkar, S.; Saha, D.; Ray, R.K. Deep Learning-Based Forecasting of Groundwater Level Trends in India: Implications for Crop Production and Drinking Water Supply. *ACS EST Eng.* **2021**, *1*, 965–977. [CrossRef]
26. Wunsch, A.; Liesch, T.; Broda, S. Groundwater Level Forecasting with Artificial Neural Networks: A Comparison of Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), and Non-Linear Autoregressive Networks with Exogenous Input (NARX). *Hydrol. Earth Syst. Sci.* **2021**, *25*, 1671–1687. [CrossRef]

27. Pietersen, K.; Beekman, H. *Groundwater Management in the Southern African Development Community*; Southern African Development Community Groundwater Management Institute: Bloemfontein, South Africa, 2016.
28. Nijsten, G.-J.; Sterckx, A.; Gomo, M.; Lukas, E. *SADC Framework for Groundwater Data Collection and Management*; Southern African Development Community Groundwater Management Institute: Bloemfontein, South Africa, 2019; p. 112.
29. Sterckx, A.; Nijsten, G.-J.; Gomo, M.; Lukas, E.; Kukuric, N. *Capacity Building for Groundwater Data Collection and Management in SADC Member States*; International Groundwater Resources Assessment Centre: Delft, The Netherlands, 2019.
30. Gaffoor, Z.; Pietersen, K.; Jovanovic, N.; Bagula, A.; Kanyerere, T. Big Data Analytics and Its Role to Support Groundwater Management in the Southern African Development Community. *Water* **2020**, *12*, 2796. [CrossRef]
31. Gibson, K. The Application of Machine Learning for Groundwater Level Prediction in the Steenkoppies Compartment of the Gauteng and North-West Dolomite Aquifer, South Africa. Master's Thesis, University of the Free State, Bloemfontein, South Africa, 2020.
32. Kanyama, Y.; Ajoodha, R.; Seyler, H.; Makondo, N.; Tutu, H. Application of Machine Learning Techniques In Forecasting Groundwater Levels in the Grootfontein Aquifer. In Proceedings of the 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Kimberley, South Africa, 25–27 November 2020; IEEE: Kimberley, South Africa, 2020; pp. 1–8.
33. Kombo, O.; Kumaran, S.; Sheikh, Y.; Bovim, A.; Jayavel, K. Long-Term Groundwater Level Prediction Model Based on Hybrid KNN-RF Technique. *Hydrology* **2020**, *7*, 59. [CrossRef]
34. Altchenko, Y.; Villholth, K.G. Transboundary Aquifer Mapping and Management in Africa: A Harmonised Approach. *Hydrogeol. J.* **2013**, *21*, 1497–1517. [CrossRef]
35. Habgood, F.; Holt, D.N.; Walshaw, R.D. *The Geology of the Thyolo Area*; Malawi Ministry of Agriculture and Natural Resources: Lilongwe, Malawi, 1973; p. 8.
36. Chairuca, L.; Chintengo, P.; Ebrahim, G.; Fraser, C.; Lautze, J.; Lazurko, A.; Macaringue, F.; Magombeyi, M.; Miranda, N.; Mokomela, R.; et al. *Transboundary Diagnostic Analysis of the Shire River Aquifer System*; Southern African Development Community Groundwater Management Institute: Bloemfontein, South Africa, 2019.
37. Rivett, M.O.; Budimir, L.; Mannix, N.; Miller, A.V.M.; Addison, M.J.; Moyo, P.; Wanangwa, G.J.; Phiri, O.L.; Songola, C.E.; Nhlema, M.; et al. Responding to Salinity in a Rural African Alluvial Valley Aquifer System: To Boldly Go beyond the World of Hand-Pumped Groundwater Supply? *Sci. Total Environ.* **2019**, *653*, 1005–1024. [CrossRef]
38. Milewski, A.M.; Thomas, M.B.; Seyoum, W.M.; Rasmussen, T.C. Spatial Downscaling of GRACE TWSA Data to Identify Spatiotemporal Groundwater Level Trends in the Upper Floridan Aquifer, Georgia, USA. *Remote Sens.* **2019**, *11*, 2756. [CrossRef]
39. Solgi, R.; Loáiciga, H.A.; Kram, M. Long Short-Term Memory Neural Network (LSTM-NN) for Aquifer Level Time Series Forecasting Using in-Situ Piezometric Observations. *J. Hydrol.* **2021**, *601*, 126800. [CrossRef]
40. Pham, Q.B.; Kumar, M.; Di Nunno, F.; Elbeltagi, A.; Granata, F.; Islam, A.R.M.T.; Talukdar, S.; Nguyen, X.C.; Ahmed, A.N.; Anh, D.T. Groundwater Level Prediction Using Machine Learning Algorithms in a Drought-Prone Area. *Neural Comput. Appl.* **2022**, *34*, 10751–10773. [CrossRef]
41. Bowes, B.D.; Sadler, J.M.; Morsy, M.M.; Behl, M.; Goodall, J.L. Forecasting Groundwater Table in a Flood Prone Coastal City with Long Short-Term Memory and Recurrent Neural Networks. *Water* **2019**, *11*, 1098. [CrossRef]
42. Kotu, V.; Deshpande, B. Classification. In *Predictive Analytics and Data Mining*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 63–163. ISBN 978-0-12-801460-8.
43. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*; Curran Associates Inc.: Long Beach, CA, USA, 2017; pp. 3149–3157.
44. Elith, J.; Leathwick, J.R.; Hastie, T. A Working Guide to Boosted Regression Trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef]
45. Kingsford, C.; Salzberg, S.L. What Are Decision Trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013. [CrossRef] [PubMed]
46. Song, Y.; Lu, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135.
47. Schapire, R.E. The Boosting Approach to Machine Learning: An Overview. In *Nonlinear Estimation and Classification*; Lecture Notes in Statistics; Springer: New York, NY, USA, 2003; Volume 171, pp. 149–171. ISBN 978-0-387-21579-2.
48. Van Houdt, G.; Mosquera, C.; Nápoles, G. A Review on the Long Short-Term Memory Model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
49. Zaini, N.; Ean, L.W.; Ahmed, A.N.; Malek, M.A. A Systematic Literature Review of Deep Learning Neural Network for Time Series Air Quality Forecasting. *Environ. Sci. Pollut. Res.* **2022**, *29*, 4958–4990. [CrossRef]
50. Smagulova, K.; James, A.P. A Survey on LSTM Memristive Neural Network Architectures and Applications. *Eur. Phys. J. Spec. Top.* **2019**, *228*, 2313–2324. [CrossRef]
51. Farzad, A.; Mashayekhi, H.; Hassanpour, H. A Comparative Performance Analysis of Different Activation Functions in LSTM Networks for Classification. *Neural Comput. Appl.* **2019**, *31*, 2507–2521. [CrossRef]
52. NASA JPL GRACE. Available online: https://grace.jpl.nasa.gov/mission/grace (accessed on 4 January 2022).
53. Swenson, S.; Wahr, J. Methods for Inferring Regional Surface-Mass Anomalies from Gravity Recovery and Climate Experiment (GRACE) Measurements of Time-Variable Gravity. *J. Geophys. Res. Solid Earth* **2002**, *107*, ETG-3. [CrossRef]
54. Tapley, B.D.; Bettadpur, S.; Watkins, M.; Reigber, C. The Gravity Recovery and Climate Experiment: Mission Overview and Early Results. *Geophys. Res. Lett.* **2004**, *31*, 4. [CrossRef]

55. Landerer, F.W.; Cooley, S.S. *Gravity Recovery and Climate Experiment Follow-on (GRACE-FO): Level-3 Data Product User Handbook*; NASA Jet Propulsion Laboratory: Pasadena, CA, USA, 2021.
56. Wahr, J.; Molenaar, M.; Bryan, F. Time Variability of the Earth's Gravity Field: Hydrological and Oceanic Effects and Their Possible Detection Using GRACE. *J. Geophys. Res. Solid Earth* **1998**, *103*, 30205–30229. [CrossRef]
57. Save, H.; Bettadpur, S.; Tapley, B.D. High-Resolution CSR GRACE RL05 Mascons. *J. Geophys. Res. Solid Earth* **2016**, *121*, 7547–7569. [CrossRef]
58. Rodell, M.; Houser, P.R.; Jambor, U.; Gottschalck, J.; Mitchell, K.; Meng, C.-J.; Arsenault, K.; Cosgrove, B.; Radakovich, J.; Bosilovich, M.; et al. The Global Land Data Assimilation System. *Bull. Am. Meteorol. Soc.* **2004**, *85*, 381–394. [CrossRef]
59. Rodell, M.; Chen, J.; Kato, H.; Famiglietti, J.S.; Nigro, J.; Wilson, C.R. Estimating Groundwater Storage Changes in the Mississippi River Basin (USA) Using GRACE. *Hydrogeol. J.* **2007**, *15*, 159–166. [CrossRef]
60. Miro, M.; Famiglietti, J. Downscaling GRACE Remote Sensing Datasets to High-Resolution Groundwater Storage Change Maps of California's Central Valley. *Remote Sens.* **2018**, *10*, 143. [CrossRef]
61. Afan, H.A.; Ibrahem Ahmed Osman, A.; Essam, Y.; Ahmed, A.N.; Huang, Y.F.; Kisi, O.; Sherif, M.; Sefelnasr, A.; Chau, K.; El-Shafie, A. Modeling the Fluctuations of Groundwater Level by Employing Ensemble Deep Learning Techniques. *Eng. Appl. Comput. Fluid Mech.* **2021**, *15*, 1420–1439. [CrossRef]
62. Guo, F.; Yang, J.; Li, H.; Li, G.; Zhang, Z. A ConvLSTM Conjunction Model for Groundwater Level Forecasting in a Karst Aquifer Considering Connectivity Characteristics. *Water* **2021**, *13*, 2759. [CrossRef]
63. Malakar, P.; Mukherjee, A.; Bhanja, S.N.; Ray, R.K.; Sarkar, S.; Zahid, A. Machine-Learning-Based Regional-Scale Groundwater Level Prediction Using GRACE. *Hydrogeol. J.* **2021**, *29*, 1027–1042. [CrossRef]