Predicting changes in spatiotemporal groundwater storage through the integration of multi-satellite data and deep learning models

JAE YOUNG SEO AND SANG-IL LEE
Department of Civil and Environmental Engineering, Dongguk University, Seoul 04620 South Korea

Corresponding author: Sang-Il Lee (e-mail: islee@dongguk.edu).

This work was supported by the National Research Foundation of Korea (NRF) grant by the Korean government (2021R1A2C011193).

ABSTRACT Continuous monitoring and accurate spatiotemporal groundwater storage change predictions can help support sustainable development and efficient groundwater resources management. In this study, remote sensing-based data were used to develop two deep learning predictive models, long short term memory (LSTM), and convolutional neural network-LSTM (CNN-LSTM) models. The input variables were terrestrial water storage anomaly (TWSA) recorded by the Gravity Recovery and Climate Experiment (GRACE) satellites and the GRACE-Follow-On (GRACE-FO) mission, precipitation measured by the Tropical Rainfall Measuring Mission (TRMM) satellite, average temperature and soil moisture obtained from the Global Land Data Assimilation System (GLDAS), and normalized difference vegetation index (NDVI) and modified normalized difference water index (MNDWI) acquired from the Landsat 5 and 8 satellites. A comparison of groundwater storage change predictions by the two deep learning predictive models with in situ measurements of the National Groundwater Monitoring Network (NGMN) in South Korea showed that the CNN-LSTM model was more accurate [root mean square error (RMSE) = 44.92 mm/month, correlation coefficient (r) = 0.70, index of agreement (IOA) = 0.81] than the LSTM model [RMSE = 47.44 mm/month, r = 0.62, and IOA = 0.77]. Furthermore, a comparison of the performance of the two models for three specific regions showed that the CNN-LSTM model captured spatiotemporal variations in groundwater storage better. Additionally, a comparison of cumulative change trends in groundwater storage with the NDVI, MNDWI, and TWSA data showed that changes in land cover and total water storage affect groundwater storage. These comparison results demonstrate that the use of parameters recorded by multiple satellites as training data for deep learning models can play an important role in the model's performance.

INDEX TERMS Groundwater storage changes, Deep learning, LSTM, CNN-LSTM, Satellite, GRACE & GRACE-FO, TRMM, GLDAS, Landsat

I. INTRODUCTION
The importance of a stable water supply and systematic water resource management has increased with the frequent occurrence of abnormal weather conditions due to climate change [1]. Groundwater is a vital freshwater resource that can help overcome these problems, and continuous and systematic groundwater change monitoring is essential for the system design, development, and implementation for effective groundwater resource management [2]. In South Korea, observation networks are operated nationwide by government departments to continuously monitor groundwater and achieve sustainable use [3], [4]. Reliable and accurate spatiotemporal groundwater storage change predictions are important for gauging the amount of groundwater available and efficiently managing groundwater resources [5]–[7]. Groundwater variations can be determined based on hydrogeological properties and boundary conditions of groundwater systems as well as meteorological, hydrological, and land cover change effects. Systematic groundwater predictions are challenging because of various nonlinearities and the complex spatial groundwater variability [5].

In recent years, remote sensing approaches have been proposed for estimating spatiotemporal groundwater storage changes, especially by combining multi-satellite data such as the Gravity Recovery and Climate Experiment (GRACE) and GRACE-Follow-On (GRACE-FO) satellites, Tropical
Rainfall Measuring Mission (TRMM), and Landsat based on the water balance method [7]–[10]. Although these approaches were evaluated relatively simply and reliably over a wide range of regions, they are not suitable for forecasting or nowcasting groundwater storage changes owing to the low spatial/temporal resolution of satellite data. Furthermore, since the GRACE satellite-based water balance method involves only hydrological components (Δterrestrial water storage, Δsnow water equivalent, Δsurface water, and Δsoil moisture, [9]), vegetation or anthropogenic effects are often ignored. Recently, some studies proved that vegetation or land cover changes including NDVI and MNDWI, are dominant predicting factors of groundwater recharge/potential [4], [11], [12]. This study suggests using multiple satellite data considering various influencing factors to increase result reliability.

The main challenge is to predict how changes in various earth system components affect groundwater storage changes, and therefore, it is necessary to analyze the groundwater response to such changes by quantifying the relationship between changes in climate factors, surface water extent, hydrological components, and vegetation [5], [6], [11].

Machine learning approaches such as those involving artificial neural networks (ANN), support vector machine (SVM), and random forest (RF) based on nonlinear interdependence have been developed to solve complicated problems in computer vision, image recognition, and time-series forecasting. Various machine learning methods have been applied to the development of groundwater prediction models, e.g., hybrid feedforward neural network (FNN) [6], [13], nonlinear autoregressive models with exogenous input (NARX) [14], [15], and ANN [7], [16] models. Although there has been an attempt to predict hydrological components, groundwater storage changes are still challenging because most studies have focused on ground observation point data. In particular, the resolution with which spatiotemporal patterns can be predicted is limited by the scale of ground observations (location, time interval, etc.), which are the input data [17]. One of the studies that attempted to overcome these limitations, Reference [7] used an ANN algorithm and multi-satellite data for monitoring spatiotemporal variations in groundwater storage and droughts in the Korean Peninsula. While this method was shown to be effective for monitoring groundwater storage variations with relatively good performance (training period and test period correlation coefficient $r \sim 0.9$ and $r \sim 0.6$, respectively), input parameter selection and overfitting in specific regions (e.g., coastal regions and data-scarce regions) are issues that need to be addressed.

In recent years, deep learning models, such as the long short term memory (LSTM) model, which is one of the recurrent neural networks (RNN) models, and the convolutional neural network (CNN) model have been developed to solve long-term dependency problems in complex and nonlinear prediction and they have shown good performance [18]. These models have multiple hidden layers that facilitate the use of complex and deep network structures to accurately identify the relationship between predictor and target variables [19]. Gradient vanishing (exploding) problems and overfitting issues faced in the case of machine learning models can be reduced using deep learning models. Furthermore, a rectified linear unit (ReLU) is an activation function that can be used to easily solve vanishing gradient problems through the weight update process with a high dropout rate [20]. Construction of predictive models by using a deep learning model typically involves five stages: 1) preprocessing for data as use input/output data, 2) designing of the model with an appropriate number of layers for defining the input/outputs, 3) model training by using loss functions, 4) network optimization (hyperparameter optimization), and 5) model validation.

LSTM and CNN are generally applied in many fields, including water resources such as terrestrial water storage [21], [22], surface water [23], [24], and flood forecasting [25], while have only recently begun to be applied to groundwater prediction in the last two years [5]. Reference [26] developed groundwater potential maps using SVM and CNN models and proved that CNN has better predictive ability than SVM. References [5], [27], [28] compared NARX, LSTM and CNN for predicting groundwater levels at the point scale of groundwater observations. Although the deep learning models improved the groundwater prediction skills, the model was evaluated at the point scale. In addition, the input data were considered only meteorological observation data (e.g., precipitation and temperature) and in situ groundwater data. Therefore, it is necessary to improve the spatiotemporal groundwater storage change prediction model considering various factors, such as meteorological, vegetation, and hydrological components.

In this study, we combined deep learning models with remote sensing-based data for various influencing factors (meteorological, land cover, and hydrological), which can provide continuous and spatial information for predicting spatiotemporal changes in groundwater storage. To our knowledge, this is the first attempt to use deep learning for groundwater predicting models based on remote sensing data in South Korea. Considering the potential effectiveness of a combination of deep learning models and remote sensing image data, this study aimed to (1) develop deep learning models–LSTM and CNN-LSTM models—for predicting groundwater storage changes from multi-satellite data and images, (2) generate long-term spatial groundwater storage changes maps for South Korea, and (3) identify groundwater storage changes based on land change.

II. MATERIALS AND METHODS
A. STUDY AREA AND GROUND OBSERVATIONS

The study area covers South Korea (Fig. 1), which is located in temperate East Asia and has a monsoonal climate. There are five major river basins (Han, Nakdong, Geum, Seomjin, and
Yeongsan) that are divided by mountain ridges, and most rivers flow westward because the eastern part is at a higher altitude than the western part (Fig. 1(a)). Shallow alluvial deposits exist along major rivers, which tend to recharge in the eastern highlands and discharge in the western lowlands. The National Groundwater Monitoring Network (NGMN) in South Korea provides information on groundwater monitoring wells operated by the National Groundwater Information Management and Service Center [29]. The are two types of monitoring wells: shallow alluvial (mean well depth: 30 m) and deep bedrock (mean well depth: 70 m depth) aquifers. Only bedrock aquifer monitoring wells have been constructed in areas without alluvial deposits [30], [31]. Based on data availability and data continuity, 166 groundwater monitoring stations (95 alluvial aquifers and 71 bedrock aquifers) were selected (Fig. 1(b)) and their groundwater levels were obtained. Monthly observations from 2003 to 2019 were used for estimating groundwater storage changes (GWSC) and for training the LSTM and CNN-LSTM models.

B. REMOTE SENSING-BASED DATA
In this study, remote sensing-based data of parameters that directly or indirectly affect groundwater storage were used as the input for the deep learning models. The inputs were precipitation (P) and average temperature (T), which were meteorological factors, normalized difference vegetation index (NDVI) and modified normalized difference water index (MNDWI), which were proxy factors for land cover, and soil moisture content (SM) and terrestrial water storage anomaly (TWSA), which were hydrological factors. Satellite products were adopted (Table 1): P from TRMM, T and SM from the Global Land Data Assimilation System (GLDAS), NDVI and MNDWI from Landsat 5 and 8, and TWSA from GRACE and GRACE-FO. The data covered the period from January 2003 to December 2019.

TABLE 1
REMOTE SENSING DATASETS USED IN THIS STUDY

| Components                        | Product          | Spatial Resolution | Period               |
|-----------------------------------|------------------|--------------------|----------------------|
| Precipitation (P)                 | TMPA3B43V7       | 0.25° × 0.25°      | 2003–2019            |
| Average temperature (T)           | GLDAS NOAH025M   | 0.25° × 0.25°      | 2003–2019            |
| Soil moisture content (SM)        | Landsat 5        | 30 m × 30 m        | 2003.01 – 2013.02    |
|                                   | Landsat 8        | 30 m × 30 m        | 2013.03 – 2019.12    |
| NDVI                              | Landsat 5        | 30 m × 30 m        | 2013.03 – 2019.12    |
| MNDWI                             | Landsat 8        | 30 m × 30 m        | 2017.06 – 2019.12    |
| Terrestrial water storage anomaly | GRACE RL06M      | 0.25° × 0.25°      | 2003.01 – 2017.06    |
| (TWSA)                            | GRACE-FO RL06M   | 0.25° × 0.25°      | 2018.06 – 2019.12    |

1) TRMM SATELLITE PRODUCT
The TRMM provides multiple satellite precipitation products (TMPA) since 1988. TMPA3B43V7 is one of the TMPAs, and it is obtained by combining TRMM microwave imager data, geosynchronous infrared estimates, and gauge data from the Global Precipitation Climatology Center (GPCC, [32]). TMPA3B43V7 provides monthly precipitation with a spatial resolution of 0.25° (approximately 25 km). Previous studies...
have evaluated the TMPA3B43 products over South Korea and have found good agreement with in situ data in terms of the correlation coefficient ($r$: 0.94–0.96) [33], [34].

2) GLDAS REANALYSIS DATA

The GLDAS is a data assimilation system based on global land surface models such as the Community Land Model (CLM), the Noah, the Mosaic, and the Variable Infiltration Capacity (VIC) models that use ground observations and multiple satellites [35]. It provides various hydrometeorological reanalysis data using multi-satellite data and ground point data, and it is used for spatial analysis and climate change and weather forecasting worldwide. Previous studies have evaluated GLDAS products such as precipitation, evapotranspiration and soil moisture over South Korea and have found good performance with ground observation data ($r$: 0.7–0.9) [36]–[39]. In this study, $T$ and the sum of the SM of each layer (0.0–2.0 m; four layers) were extracted from the GLDAS Noah model at monthly intervals with a spatial resolution of 0.25° × 0.25°.

3) LANDSAT SATELLITE-BASED LAND COVER INDEX

Landsat imagery has been widely used to study and monitor various aspects of inland cover [40]. Landsat satellites provide global coverage with a resolution of 30 m, and Level-1 T1 images covering South Korea are available from the USGS Earth Explorer [41]. A total of 2,652 images were selected from the Landsat 5 Thematic Mapper (TM) for the period 2003.01–2013.02 and from the Landsat 8 Operational Land Imager (OLI) for the period 2013.03–2019.12. Radiometric calibration and atmospheric correction to remove clouds and shadows were performed using the fast line-of-sight atmospheric analysis of spectral hypercubes (FLAASH) algorithm based on metadata files from the USGS [42]. Since there were missing data after calibration, Landsat images were monthly averaged to generate NDVI and MNDWI time series and mosaiced using nearest-neighbor mean resampling over the study area. NDVI was calculated as shown in (1) for all images by using the near-infrared (NIR) band and red visible band (RED) of calibrated Landsat images, and MNDWI was computed using the green visible band (GREEN) and the shortwave infrared (SWIR) band, as shown in (2). The NDVI and MNDWI values range from -1 to 1, and high NDVI and MNDWI values reflect healthy vegetation and water bodies, respectively.

\[ \text{NDVI} = \frac{(\text{NIR} – \text{RED})}{(\text{NIR} + \text{RED})} \quad (1) \]
\[ \text{MNDWI} = \frac{(\text{GREEN} – \text{SWIR})}{(\text{GREEN} + \text{SWIR})} \quad (2) \]

NDVI and MNDWI with a 30 m spatial resolution were converted to fractional values at 0.25° to match their spatial resolution with that of the other input variables. For fractional value extraction, thresholds were applied to vegetation and non-vegetation regions, and water and non-water features were segmented by considering both indices. The fractions of vegetation and water cover were calculated using the following equation:

\[ f_{\text{veg}} = \frac{\text{pixels}(\text{veg})}{\text{pixels}(\text{veg} + \text{nonveg})} \quad (3) \]
\[ f_{\text{water}} = \frac{\text{pixels}(\text{water})}{\text{pixels}(\text{water} + \text{nonwater})} \quad (4) \]

where $f_{\text{veg}}$ and $f_{\text{water}}$ are fractional vegetation and water cover, respectively, and pixels(veg), pixels(nonveg), pixels(water), and pixels(nonwater) are the number of vegetation, non-vegetation, water and non-water pixels, respectively.

References [43] and [44] determined the maximum NDVI threshold for bare soil and vegetation over South Korea as 0.31 and 0.86 by using Landsat images, respectively. Furthermore, many previous studies have set the MNDWI threshold value for extracting water bodies to zero. In this study, the threshold was set to be in the range of 0.31–0.86 for vegetation and 0 for water bodies.

4) GRACE AND GRACE-FO SATELLITE MASCON SOLUTION

The GRACE mission, launched by NASA and Deutsches Zentrum für Luft-und Raumfahrt (DLR), was operated between 2002 and 2017. GRACE data include temporary gravitational field changes due to mass changes in the atmosphere, oceans, water, and ice zones. The success of the GRACE mission led to the launch of GRACE-FO in May 2018 for the same purpose [45].

Changes in the gravitational field can be converted into changes in the equivalent water thickness or terrestrial water storage (TWS). GRACE/GRACE-FO Level-3 (RL06) data include TWSA data processed using methods, which can be determined in two possible ways by using either spherical harmonic coefficients or a mass concentration function (Mascon). GRACE/GRACE-FO data are available from various institutes, including the University of Texas at Austin Center for Space Research (CSR), the GeoForschungsZentrum (GFZ), and the Jet Propulsion Laboratory (JPL). In this study, TWSA was extracted from CSR RL06M (RL06 Mascon), the latest data of GRACE and GRACE-FO [46], [47]. CSR RL06M TWSA data were provided in a 0.25° × 0.25° grid, but they represent hexagonal tiles with an equal-area geodesic grid size of 1.0° × 1.0° at the equator [47]. There were temporal gaps between GRACE and GRACE-FO with a length of approximately 14 months. Many studies have reported on the gap filling between GRACE and GRACE-FO temporal data using interpolation (linear [48], singular spectrum analysis [48], and cubic spline [49], [50]). Reference [48] applied linear interpolation to fill the gap and confirmed that the trend is positively biased by over 10% compared to the original series. The interpolation results of different studies may vary, and the original series will also affect the prediction. Nevertheless, since no data can replace the GRACE and GRACE-FO data, the interpolation method is widely used. In this study, a cubic spline suitable for time series interpolation was applied [49], [50].
C. FRAMEWORK OF DEEP LEARNING MODEL

1) MODEL STRUCTURE

The deep learning models employed in this study were LSTM and CNN-LSTM models. LSTM is one of the deep learning algorithms that are commonly used for regression, especially for time-series prediction, and it is also used for solving the vanishing gradient problem for long-term predictions [18]. LSTM is slightly more complex than the traditional RNN when calculating hidden states and adds cell memory. In the LSTM cell memory, three types of gates (input gates, forget gates, and output gates) delete unnecessary memories and determine what should be remembered by the network. LSTM has a conveyor-belt-like structure that contains cell states, and each cell state has a learner to update the state.

For an input sequential set of \( x = [x_1, x_2, \cdots, x_n] \) and an output set, the process is expressed below. The input gate \( i_t \) is the gate for remembering the current information, and \( g_t \) memorizes the updated information as follows:

\[
i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{5}
\]

\[
g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \tag{6}
\]

where \( \sigma \) is the activation function, \( W \) is the weight, \( h \) is the hidden state, and \( b \) is the bias term.

The forget gate \( f_t \) deletes memories, and the \( x_t \) and the hidden state at \( t-1 \) (\( h_{t-1} \)) pass through the activation function. The input gate and forget gate determine the new memory cell state (\( C_t \)) by an elementwise product (\( \odot \)) as follows:

\[
f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{7}
\]

\[
c_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{8}
\]

Finally, the output gate \( O_t \) is used to determine the hidden state at \( t \) (\( h_t \)). The value of the long-term state passes through the hyperbolic tangent function and becomes a value between -1 and 1. Hidden state values are directed to the output layer.

\[
O_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{9}
\]

\[
h_t = O_t \odot \tanh(C_t) \tag{10}
\]

A CNN is advantageous for feature extraction and consists of convolution layers and max pooling layers. Each convolution layer contains convolution kernels. CNN-LSTM is a CNN combined with the LSTM architecture, designed specifically for the problem of sequence prediction of spatial input data such as images. CNN was used for the spatial information extraction from satellite images or data and LSTM was used for sequence prediction in the CNN-LSTM model. In other words, after extracting features between several input variables on CNN, time information on irregular trends in time-series components was modeled in LSTM based on the extracted information.

In this study, remote sensing data-driven LSTM and CNN-LSTM models were compared and analyzed. The models were constructed using TensorFlow in the Keras library of Python (version 3.6), and Fig. 2 shows the LSTM and CNN-LSTM structures. The LSTM model comprised an input layer, and the LSTM layer consisted of each cell memory, a dense layer fully

---

**FIGURE 2.** Structure of deep learning models used to predict GWSC: (a) LSTM and (b) CNN-LSTM
connected to hidden nodes, and an output layer. In the CNN-LSTM model, data are sent from the input layer to the 2D convolution layer of the CNN and further transmitted to the LSTM model through the max-pooling layer and flatten layer.

The remote sensing-based input variables were normalized using the min-max scaling method to avoid disrupting convergence in the network. This method is defined by (11), which can be transformed to confine the data to the range 0–1.

\[ x'_i = \frac{(x_i - \min(x_i))}{\max(x_i) - \min(x_i)} \quad (11) \]

The ReLU function was adopted for use as an activation function since it can converge faster than the sigmoid or tangent hyperbolic function. The loss function for updating the network was selected as the mean squared error (MSE) coupled with the Adam optimizer. The Adam optimizer is a commonly used technique for determining weights for minimizing the loss function in the task of making time-series predictions.

2) HYPERPARAMETER TUNING

Hyperparameters are variables used in deep learning models, and they are generally based on the prior knowledge of the experimenter about the optimal implementation of deep learning models [51]. Deep learning is a complex type of neural network with various hyperparameters such as batch size, training iterations, and hidden layers that should be tuned before training. Hyperparameter tuning improves the performance of deep learning models and helps avoid overfitting or underfitting problems. Hyperparameter tuning techniques include manual search, grid search, random search, and Bayesian optimization. Bayesian optimization is a method of finding hyperparameters by using Bayes rules on the basis of the experimenter’s prior knowledge and it takes a relatively shorter time to determine optimal parameters for large datasets compared with grid search and random search [52].

In this study, Bayesian optimization methods based on Python’s Keras tuning library were used for cross-validation to determine the optimal hyperparameters. Hyperparameter sets and the optimization range of values were configured as shown in Table 2. The time series were divided into training and test periods, and these two periods were set to 60% (January 2003 to February 2013) and 40% (March 2013 to December 2019) of the total period, respectively.

| Model   | Hyperparameter     | Optimization range of values |
|---------|--------------------|------------------------------|
| LSTM    | Number of hidden layers | {1, 2, ..., 5}               |
|         | Number of nodes per layer | {5, 10, ..., 100}           |
|         | Time lag (month)    | {1, 2, ..., 6}               |
|         | Dropout rate        | {0.1, 0.2, ..., 0.5}        |
|         | Batch size          | {5, 10, ..., 100}           |
|         | Epochs              | {50, 100, ..., 500}         |
| CNN-LSTM| Number of filters   | {16, 32, ..., 128}          |
|         | Number of hidden layers | {1, 2, 3}                  |
|         | Number of nodes per layer | {5, 10, ..., 100}           |
|         | Time lag (month)    | {1, 2, ..., 6}               |
|         | Dropout rate        | {0.1, 0.2, ..., 0.5}        |
|         | Batch size          | {5, 10, ..., 100}           |
|         | Epochs              | {50, 100, ..., 500}         |

III. RESULTS

A. MODEL PERFORMANCE DEPENDENCE ON HYPERPARAMETER OPTIMIZATION

In the deep learning modeling process, training the model by setting the hyperparameters has a significant influence on the model performance. The purpose of training is to update parameters to minimize loss values through optimization methods [25]. Fig. 3 shows the loss results of both the LSTM and CNN-LSTM models for Bayesian optimization for each latency and the number of hidden layers. The LSTM model had a greater loss for 1 and 2 months of delay compared with a lag time of 3–6. For a lag time of more than three months, the more hidden layers there were, the greater the loss value. The CNN-LSTM model performance improved when the lag time increased, and converged from six months. The final network architecture of the LSTM and CNN-LSTM models for GWSC prediction was designed to have one hidden layer with a lag time of three months and two hidden layers with a lag time of six months.

The GWSC data predicted by each model were compared with observations at ground stations distributed across South Korea. Groundwater level change data (\( \Delta W \)) from groundwater stations were converted to GWSC in terms of equivalent height by using a specific yield (\( Y_s \)) (12). In this study, the specific yield was set to 0.20, which was reported in previous studies [7], [8].

\[
GWSC = \Delta W \times Y_s \quad (12)
\]

The test results (March 2013 to December 2019) were examined in terms of the root mean square error (RMSE),
correlation coefficient \((r)\), and index of agreement (IOA) between the observed and predicted GWSCs to assess the accuracies of the two models.

Fig. 4 shows the density scatterplots showing the correlation between all cells of the predicted GWSC and in situ measurements corresponding to locations of the cells during the test period. The GWSC predicted by the CNN-LSTM model showed a higher correlation with in situ measurements (Fig. 4(b)) and higher accuracy (RMSE = 44.92 mm/month, \(r = 0.70\), and IOA = 0.81) than that predicted by the LSTM model during the test period (March 2013 to December 2019). While the LSTM model (RMSE = 47.44 mm/month, \(r = 0.62\), and IOA = 0.77) was less accurate (Fig. 4(a)), its result did not differ significantly from the CNN-LSTM model’s result.

**B. EVALUATION OF PREDICTED GROUNDWATER STORAGE CHANGES**

The deep learning models performance over South Korea is shown by the time series in Fig. 5. The magnitude and variability of GWSCs obtained from the LSTM (red dotted line; Fig. 5(a)) and CNN-LSTM (green line; Fig. 5(b)) models were consistent with those of in situ measurements (black line). The TMPA precipitation during the study period is also presented in Fig. 5. The LSTM and CNN-LSTM models captured both the monthly and seasonal dynamics of in situ GWSC associated with precipitation, apart from showing stable performance for the test period (March 2013 to December 2019). However, a comparison of the LSTM model’s predictions with in situ measurements indicated that the model tended to overestimate GWSCs.

Fig. 6 shows the spatial distribution maps for the RMSE, correlation coefficient, and IOA obtained by comparing grid results from deep learning models and 0.25° spatial resolution of in situ measurements by using inverse distance weighting (IDW) during the test period (March 2013–December 2019). The grids on the red or yellow scale are characterized by large errors and negative or non-significant correlation, respectively, whereas the blue scale represents cells with significant linear correlations. Comparisons between deep learning model predictions and in situ measurement showed good agreement overall, with only small areas being characterized by high RMSE values. Overall, the CNN-LSTM model’s predictions appeared to be better than the LSTM model’s predictions, and for both models, higher RMSE values were notably observed in the southwest coastal and eastern mountainous regions. This error appears to result from the limited number of training samples for these areas.
FIGURE 5. GWSC time-series comparisons predicted by the LSTM (red dotted line in (a)), and CNN-LSTM (green line in (b)) models with in situ measurements (black line) corresponding to TMPA precipitation.

FIGURE 6. Spatial distribution of the RMSE, correlation coefficient ($r$), and IOA between deep learning model-predicted GWSC and interpolated in situ data during the test period (March 2013-December 2019).
C. SPATIOTEMPORAL GROUNDWATER STORAGE CHANGES AND LAND COVER CHANGES

Monthly GWSC maps generated for South Korea were used to visually evaluate the performance of the proposed deep learning models through the test period results. The interpolated in situ data were compared with the values obtained from the two proposed deep learning models. Fig. 7 shows three examples of GWSC maps obtained with the predictions of the CNN-LSTM and LSTM models as well as those obtained from in situ measurements for March 2019 and August 2019.

Overall, the GWSC spatial patterns for all selected examples (March and August in 2019) were well predicted by the LSTM and CNN-LSTM models (second and third columns in Fig. 7), as evidenced from a comparison with the corresponding maps generated from the in situ data (first column in Fig. 7).

The CNN-LSTM model simulated the spatial patterns of GWSCs fairly accurately. The high variability in the eastern (blue areas shown for March 2019) and northern and southeastern areas (blue and red areas for August 2019) were well matched. However, for the LSTM model, although the spatial variability pattern of the GWSC appeared to be in good agreement, the predicted maps could not capture the variability as the maps derived from the CNN-LSTM predictions and in situ measurements did. This showed the superiority of the CNN-LSTM model, which was predominantly attributed to its convolutional filter.

The GWSC, land cover index and TWSA were compared to analyze whether the influencing factors considered in the deep learning model affected the GWSCs. Land cover changes (NDVI and MNDWI) recorded by Landsat 5 and 8, and the TWSA of GRACE and GRACE-FO were analyzed by comparing cumulative GWSC (in situ and deep learning predictions) from 2003.

Seoul (A in Fig. 8), a part of Jeollanam-do (B in Fig. 8), and a part of Gyeongsangbuk-do (C in Fig. 8) were chosen as regions that use groundwater the most, and they were examined in detail. A region is South Korea’s capital and the most densely populated city, and groundwater usage is 6.2% of the total amount of groundwater used in the country. The groundwater usage is 14.3% and 14.0% in regions B and C, respectively [29]. In all three regions, both the vegetation cover fraction and waterbody cover fraction decreased in 2008 compared with 2003 (Table 3 and Fig. 8), apparently because of the drought that occurred in Korea in 2008 (precipitation in

FIGURE 7. Comparison of monthly spatial distribution between interpolated GWSC and predicted GWSC data for March 2019 and August 2019 during the test period.
Fig. 5). In 2015 and 2019, NDVI and MNDWI recovered and showed fraction values similar to those for 2003. From 2003 to 2019, the annual rate of change in the vegetation cover fraction showed an increasing trend in all three regions (0.13%–2.74%/year), and the waterbody cover fraction showed a decreasing trend from -0.62%/year to -0.03%/year (Table 3).

**TABLE 3 LAND COVER CHANGE IN THE CHOSEN REGIONS FROM 2003 TO 2019**

| Time   | Vegetation fraction (%) | Waterbody fraction (%) |
|--------|-------------------------|------------------------|
|        | A   | B   | C     | A | B   | C |
| 2003.05 | 61.12 | 82.20 | 86.45 | 4.59 | 2.47 | 2.10 |
| 2008.05 | 57.20 | 79.69 | 83.20 | 3.83 | 2.11 | 0.98 |
| 2015.05 | 62.16 | 79.68 | 83.58 | 3.31 | 2.54 | 2.33 |
| 2019.05 | 64.14 | 79.04 | 86.87 | 2.56 | 2.71 | 3.28 |

![FIGURE 8. Spatial distribution of land cover change. Subsets A, B, and C are specific regions showing the changes in vegetation and waterbody cover fractions.](image)

**Fig. 8.** Spatial distribution of land cover change. Subsets A, B, and C are specific regions showing the changes in vegetation and waterbody cover fractions.

Fig. 9 shows the time series of cumulative GWSCs obtained from in situ measurements and the LSTM and CNN-LSTM models, and a linear trend for the A (Fig. 9(a)), B (Fig. 9(b)), and C (Fig. 9(c)) regions is evident. The CNN-LSTM model (green lines in Fig. 9) not only captured the decreasing trend of in situ GWSCs well (black lines in Fig. 9), but also clearly showed the highest accuracies for the three regions ($r = 0.88, 0.92, \text{and } 0.82$ for the A, B, and C regions, respectively). The LSTM model ($r = 0.87, 0.87, \text{and } 0.81$ for the A, B, and C regions, respectively) was less accurate (red dotted lines in Fig. 9). The TWSA data of GRACE and GRACE-FO had a direct influence on the hydrological dynamics, including GWSCs. Cumulative GWSCs were compared with the TWSA. The GWSC rate was inferred to be $-9.09 \pm 0.4$ mm/year for 2003.01–2019.12 in region A, similar to the rate of decrease in the TWSA estimated from GRACE and GRACE-FO data ($-5.21 \pm 0.2$ mm/year; Fig. 9(a)). In region B, the rate of depletion GWSCs was $-4.00 \pm 0.2$ mm/year (Fig. 9(b)), and the decrease in the TWSA rate was $-3.24 \pm 0.2$ mm/year, comparable to the depletion rate. During the study period, GWSCs in the C region also decreased slightly, with a rate of $-4.14 \pm 0.2$ mm/year, and TWSA showed a decrease rate of $-8.68 \pm 0.3$ mm/year.

These results indicate that the predicted GWSCs are related to influencing factors of hydrological and land cover changes from satellites. Despite the difficulty in evaluating the complex variability of groundwater mechanics using various variables, the proposed prediction algorithm based on multi-satellite data is expected to provide a new direction for the development of high-performance GWSC prediction models in a relatively simple way.

**IV. CONCLUSIONS**

Monitoring GWSC is a nonlinear and complex task, and a large amount of data for various variables and a high proficiency of researchers are required for physical modeling.
Deep learning models can provide a promising alternative to capture complex interactions between variables related to GWSC. Furthermore, satellite data for various variables can be easily incorporated to predict spatiotemporal GWSC in deep learning models.

In this study, LSTM and CNN-LSTM models were used to accurately predict GWSC over South Korea by using different feature combinations from multi-satellite data. The deep learning models involved various input variables (precipitation, average temperature, soil moisture storage changes, TWSA, NDVI, and MNDWI). It is necessary to understand the sensitivity of their predictive performance and avoid overfitting problems. Hyperparameter optimization based on Bayesian optimization was performed using different parameter combinations and the prediction accuracies of the combinations were compared.

Overall, the results showed that the CNN-LSTM model predicted GWSC maps relatively better than the LSTM model in terms of the RMSE, $r$, and IOA. The LSTM model performed rather well in shallow neural networks (one hidden layer), and when using spatiotemporal data such as satellite data, the CNN-LSTM model with an added convolution layer appeared to be better. In the LSTM model, the optimal lag time was three months and in the case of the CNN-LSTM model,
the optimal lag time was six months. The latter model showed better performance in a longer lag time than the former model.

These results indicate that the deep learning models based on satellite data proposed in this study (with meteorological, hydrological, and land cover change indicators) are suitable for long-term groundwater change modeling. It is concluded that considering influencing factors such as precipitation, land cover changes, and TWSC resulted in the models accurately predicting spatiotemporal changes in groundwater storage. Predictions for some areas with a relatively low accuracy may be improved by considering additional input data related to GWSCs or other hyperparameter optimization methods.

Furthermore, relatively accurate predictions of GWSCs, such as those provided by the proposed models can be used to determine appropriate measures to deal with climate change, thereby rendering society competent to respond to the challenges posed by climate change.

ACKNOWLEDGEMENTS

The authors would like to thank four anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] T. Gleson, M. Nagel, and Y. Wada, “Assessing regional groundwater stress for nations using multiple data sources with the groundwater footprint,” Environ. Res. Lett., vol. 8, pp. 44010, Oct. 2013.
[2] C.F. Taylor, and W.M. Alley, “Ground-water-level monitoring and the importance of long-term water-level data,” U.S. Geological Survey Circular 1217, Jan. 2002. [Online]. Available at: https://pubs.usgs.gov/circ/circ1217/pdf/circular1217.pdf
[3] V.P. Pandey, S. Shrestha, S.K. Chapagain, and F. Kuzama, “A framework for measuring groundwater sustainability,” Environ. Sci. Policy., vol. 14, no. 4, pp. 396–407, Apr. 2011.
[4] E. Parizi, S.M. Hosseini, B. Ataei-Ashtiani, and C.T. Simmons, “Normalized difference vegetation index as the dominant predicting factor of groundwater recharge in phreatic aquifers: case studies across Iran,” Sci. Rep., vol. 10, pp. 17473, Oct. 2020.
[5] A. Wunsch, T. Lesch., and S. Broda, “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),” Hydro. Earth Syst. Sci., vol. 25, pp. 1671–2021, Apr. 2021.
[6] S. Sahoo, T.A., Russo, J. Elliott, and J. Foster, “Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S,” Water Resour. Res., vol. 53, pp. 3878–3895, Apr. 2017.
[7] J.Y. Seo, and S.-I. Lee, “Spatio-temporal groundwater drought monitoring using multi-satellite data based on an artificial neural network,” Water, vol. 11, pp. 1953, Sep. 2019.
[8] J.Y. Seo, and S.-I. Lee, “Integration of GRACE, ground observation, and land-surface models for groundwater storage variations in South Korea,” Int. J. Remote Sens., vol. 37, no. 24, pp. 5786–5801, Nov. 2016.
[9] M. Rodell, I. Velicogna, and J.S. Famiglietti, “Satellite-based estimates of groundwater depletion in India,” Nature, vol. 460, pp. 999–1002, Aug. 2009.
[10] A. Hassan, and S. Jin, “Water storage changes and balances in Africa observed by GRACE and hydrologic models,” Geod. Geodyn., vol. 7, no. 1, pp. 39–49, Jan. 2016.
[11] J.H. Kim and R.B Jackson, “A global analysis of groundwater recharge for vegetation, climate, and soils,” Vadose Ze. J., vol. 11, pp. 1–36, Feb. 2012.
[12] B. A. Yifru, I.-M. Chung, M.-G. Kim, and S. W. Chang, “Assessing the effect of land/sea land cover and climate change on water yield and groundwater recharge in East African Rift Valley using integrated model,” J. Hydrol. Reg. Stud., vol. 37, pp. 100926, Oct. 2021.
[13] S. Mohanty, M.K. Jha, S.K. Raul, R.K. Panda, K.P. and Sudheer, “Using artificial neural network approach for simultaneous forecasting of weekly groundwater levels at multiple sites,” Water Resour. Manag., vol. 29, pp. 5521–5532, Oct. 2015.
[14] S.M. Guzman, J.O. Paz, and M.L.M. Tagert, “The use of NARX neural networks to forecast daily groundwater levels,” Water Resour. Manag., vol. 31, pp. 1591–1603, Mar. 2017.
[15] A. A. Alsumaieiei, “A nonlinear autoregressive modeling approach for forecasting groundwater level fluctuation in urban aquifers,” Water, vol. 12, pp. 820, Mar. 2020.
[16] A.Y. Sun, “Predicting groundwater level changes using GRACE data,” Water Resour. Res., vol. 49, pp. 1–13, Sep. 2013.
[17] X.X. Zhu, D. Tuia, L. Mou, G-S. Xia, L. Zhang, F. Xu, and J. Frankefjord, “Deep learning in remote sensing: a comprehensive review and list of resources,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2017.
[18] Z. Zhang, Y. Zhu, X. Zhang, M. Ye, and J. Yang, “Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural area,” J. Hydrol., vol. 561, pp. 918–929, Apr. 2018.
[19] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonards, J. Piater, A.J. Rodriguez-Sanchez, and L. Wiskott, “Deep hierarchies in the primate visual cortex: what can we learn for computer vision?,” IEEE Trans. Pattern Anal. Machine Intel., vol. 35, no. 8, pp. 1847–1871, Aug. 2013.
[20] L. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: surpassing human-level performance on ImageNet classification,” Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1026–1034.
[21] Q. Yu, S. Wang, H. He, K. Yang, L. Ma, and J. Li, “Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model (“, Int. J. Appl. Earth., vol. 102, pp. 102404, Oct. 2021.
[22] A.Y. Sun, B.R. Scanlon, Z. Zhang, D. Walling, S.N. Bhanja, A. Mukherjee, and Z. Zhong, “Combining physically-based modeling and deep learning for fusing GRACE satellite data: Can we learn from mismatch?,” Water Resour. Res., vol. 55, no. 2, pp. 1179–1195, Feb. 2019.
[23] S. Ghimire, Z.M. Yaseen, A.A. Farooque, R.C. Deo, J. Zhang, and X. Tao, “Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks,” Sci. Rep., vol. 11, pp. 17497, Sep. 2021.
[24] D. Liu, W. Jiang, L. Mu and S. Wang, “Streamflow prediction using deep learning neural network: Case study of Yangtze River,” IEEE Access, vol. 8, pp. 90069–90086, May 2020.
[25] M. Panahi, N. Sadhasivam, H.R. Pourghasemi, F. Rezaie, and S. Lee, “Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR),” J. Hydrol., vol. 588, pp. 125033, Sep. 2020.
[26] T. Lahivara, A. Malehm, A. Pasanen, L. Kärkkäinen, J.M.J. Huttunen, and J.S. Hesthaven, “Estimation of groundwater storage from seismic data using deep learning,” Geophys. Prospect., vol. 67, pp. 2115–2126, Jun. 2019.
[27] J. Müller, J. Park, R. Sahu, C. Varadharajan, B. Arora, B. Faybishenko, and D. Agarwal, “Surrogate optimization of deep neural networks for groundwater predictions,” J. Glob. Optim., vol. 81, pp. 203–231, May. 2020.
[28] H. Afzaal, A.A. Farooque, F. Abbas, B. Acharya, and T. Essau, “Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning,” Water, vol. 12, no. 1, pp. 5, Dec. 2019.
[29] GIMS. 2021. [Online] Available at: https://gims.go.kr, Accessed on: 9 September 2021.
[30] J.-Y. Lee, and K.D. Kwon, “Current status of groundwater monitoring networks in Korea,” Water, vol. 8, no. 4, pp. 168, Apr. 2016.
[31] J.M. Lee, N.C. Woo, C.-J. Lee, and K. Yoo, “Characterising bedrock aquifer systems in Korea using paired water-level monitoring data,” Water, vol. 9, no. 6, pp. 420, Jan. 2017.
[32] G.J. Huffman, R.F. Adler, D.T. Bolvin, G. J. Gu, E.J. Nelkin, K.P. Bowman, Y. Hong, E.F. Stocker, and D.B. Wolff, “The TRMM
multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales,” J. Hydrometeorol., vol. 8, no. 1, pp. 38–55, Feb. 2007.

[33] M.S. Koo, S.Y. Hong, and J. Kim, “An evaluation of the tropical rainfall measuring mission (TRMM) multi-satellite precipitation analysis (TMPA) data over South Korea,” Asia-Pac. J. Atmos. Sci., vol. 45, no. 3, pp. 265–282, Aug. 2009.

[34] J.Y. Seo, and S.-I. Lee, “Total discharge estimation in the Korean Peninsula using multi-satellite products,” Water, vol. 9, pp. 332, Jul. 2017.

[35] M. Rodell et al., “The Global Land Data Assimilation System,” Bull. Am. Meteor. Soc., vol. 85, no. 3, pp. 381–394, Mar. 2004.

[36] J. Baik, J. Jeong, J. Park, and M. Choi, “A study on the analyzing of uncertainty for actual evapotranspiration: flux tower, satellite-based and reanalysis based dataset,” J. Korea Water Resour. Assoc., vol. 52, no. 1, pp. 11–19, Jan. 2019 (in Korean).

[37] J. Baik and M. Choi, “Multi-satellite-based water budget components in South Korea,” Environ. Earth Sci., vol. 77, no. 3, pp. 1–19, Feb. 2018.

[38] S. Park, S. Park, J. Im, J. Rhee, J. Shin, and J.D. Park, “Downscaling GLDAS soil moisture data in East Asia through fusion of multi-sensors by optimizing modified regression trees,” Water, vol. 9, no. 5, pp. 332, May. 2017.

[39] C.S. Lee, and J.D. Park, “Improvement of AMSR2 soil moisture products over South Korea,” IEEE J. Sel. Top. Appl., vol. 10, no. 9, pp. 3839–3849, Sep. 2017.

[40] J. Yang, and X. Du, “An enhanced water index in extracting water bodies from Landsat TM imagery,” Ann. GIS, vol. 23, no. 3, pp. 141–148, Jun. 2017.

[41] USGS Earth Explorer. 2021. [Online] Available at: https://earthexplorer.usgs.gov.kr, Accessed on: 9 September 2021.

[42] T. Cooley et al., “FLAASH, a MODTRAN4-based atmospheric correction algorithm, its application and validation,” IEEE International Geoscience and Remote Sensing Symposium, 2002, pp. 1414–1418 vol.3, doi: 10.1109/IGARSS.2002.1026134.

[43] S. Choi, S. Lee, and W. Baio, “Analysis of vegetation cover fraction on Landsat OLI using NDVI,” J. Korean Soc. Surv. Geodesy Photogramm. Cartogr., vol. 32, no. 1, pp. 9–17, Feb. 2014 (in Korean).

[44] S.K. Choi, S.K. Lee, S.H. Jung, J.W. Choi, D.Y. Choi, and S.J. Chun, “Estimation of fractional vegetation cover in sand dunes using multispectral images from fixed-wing UAV,” J. Korean Soc. Surv. Geodesy Photogramm. Cartogr., vol. 34, no. 4, pp. 431–441, Aug. 2016.

[45] GRACE & GRACE-FO. 2021. [Online] Available at: https://grace.jpl.nasa.gov, Accessed on: 9 September 2021.

[46] H. Save, S. Bettpapur, and B.D. Tapley, High resolution CSR GRACE RL05 mascons, J. Geophys. Res. Solid Earth, vol. 121, no. 10, pp. 7547–7569, Sep. 2016.

[47] H. Save, “CSGR GRACE and GRACE-FO RL06 Mascon Solutions v02”, 2021. [Online] Available at: 10.15781/crg9-hl24, Accessed on: 9 September 2021.

[48] S. Yi, and N. Sneeuw, “Filling the data gaps within GRACE missions using singular spectrum analysis,” J. Geophys. Res. Solid Earth, vol. 126, no. 5, pp. e2020JB021227, May. 2021.

[49] Z.M. Nigatu, D. Fan, W. You, and A.M. Melesse, “Hydroclimatic extremes evaluation using GRACE/GRACE-FO and multidecadal climatic variables over the Nile river basin,” Remote Sens., vol. 13, pp. 651, Feb. 2021.

[50] Z. Shi, W. Zheng, and W. Yin, “Improving the reliability of the prediction of terrestrial water storage in Yunnan using the artificial neural network selective joint prediction model,” IEEE Access, vol. 9, pp. 31865–13879, Jan. 2021.

[51] R.K. Sahu, J. Muller, J. Park, C. Varadharajan, B. Arora, B. Faybishenko, and D. Agarwal, “Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network,” Front. Water, vol. 2, pp. 573034, Nov. 2020.

[52] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, Hyperparameter optimization for machine learning models based on bayesian optimization, J. Electron. Sci. Technol., vol. 17, no. 1, pp. 26–40, Mar. 2019.