Integrating GRACE/GRACE Follow-on and wells data to detect groundwater storage recovery at a small-scale in Beijing using deep learning

Ying Hu [1], Nengfang Chao [1,*], Yong Yang [2], Jiangyuan Wang [1], Wenjie Yin [3], Jingkai Xie [4], Guangyao Duan [2], Menglin Zhang [2], Xuewen Wan [1], Fupeng Li [5], Zhengtao Wang [6], Guichong Ouyang [1]

[1] College of Marine Science and Technology, Hubei Key Laboratory of Marine Geological Resources, Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences, Wuhan 430074, China

[2] Beijing Water Science and Technology Institute, Beijing 100048, China

[3] Satellite Application Center for Ecology and Environment, Ministry of Ecology and Environment, Beijing 100094, China

[4] Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576, Singapore

[5] Institute of Geodesy and Geoinformation, University of Bonn, 53115 Bonn, Germany

[6] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

* Correspondence: chaonf@cug.edu.cn

1. GRACE/GRACE-FO SH product post- preprocessing

The GRACE SH Level 2 datasets were post-preprocessed considering the following factors: 1) The degree one coefficient was replaced by the geocentric correction term calculated by Swenson et al., (2008) [1], and the original C_{20} coefficients were replaced by SRL observations [2], as the influence of the satellite's near-polar orbit design, GRACE cannot directly observe geocentric motion and is insensitive to C_{20} (degree 2 order 0) coefficients. In this study; 2) A 300 km Gaussian filter was applied to the SH coefficients to suppress high-frequency noise. [3]; 3) Long-term gravity variations caused by the process of glacial isostatic adjustment (GIA) were compensated using a GIA model [4]; 4) Signal leakage and bias were recovered using the scaling factor method based on the GLDAS data [5], [6].

2. Deep Learning

2.1 LSTM

Hochreiter and Schmidhuber, (1997) first proposed the Long Short-Term Memory Network (LSTM) [7], which is a special type of Recurrent Neural Network (RNN), effectively solving the problems of gradient explosion and vanishing gradient of RNN, and has been widely in fields such as hydrology and climate simulation. The core part of LSTM consists of three gate structures (i.e., Forget Gate, Input Gate, and Output Gate), as well as cell candidate states (as shown in Figure S1), which controlled information input and output. At the time step t, LSTM has three input variables: cell state C_{t-1}, hidden state h_{t-1}, and input features x_t, and two output variables: cell state C_t and hidden state h_t. The main function of the forget gate is to determine which information from input features

x_t and hidden state h_{t-1} should be used in calculating the cell state C_t, and which information should be forgotten. The function of the input gate is to update the cell state. First, the previous hidden state h_{t-1} and the current input features x_t are passed to the activation function σ to determine what information is retained; next, the h_{t-1} and x_t are passed to an activation function tanh to create a cell candidate state $\tilde{C}_t$; finally, multiply the output values of the two steps above to determine which information is updated. The cell state was used to determine the retained information in the cell state. The previous cell state C_{t-1} and forget vector f_t multiplied element-wise, and the information in the new cell state that is close to 0 is discarded. And then, the value is added point by point with the output value of the input gate to update the new information found by the neural network to the cell state, and finally, the updated cell state is obtained. The output gate is used to determine the hidden state h_t. The hidden state contains the previous input information. First, the previous hidden state h_{t-1} and the current input feature x_t are passed to the activation function σ, then the cell state C_t is passed to the activation function tanh, and finally, the output values of the two are multiplied to determine the information that the hidden state h_t should carry, and the hidden state h_t is output. The LSTM gate structure and the cell state are calculated as follows:

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f) \tag{S1}$$

$$\begin{aligned} i_t &= \sigma(W_i x_t + R_i h_{t-1} + b_i) \\ \tilde{C}_t &= \tanh(W_c \bullet x_t + R_o h_{t-1} + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \end{aligned} \tag{S2}$$

$$\begin{aligned} o_t &= \sigma(W_o x_t + R_o h_{t-1} + b_o) \\ h_t &= o_t \odot \tanh(C_t) \end{aligned} \tag{S3}$$

where t stands for the time step, $\odot$ is the Hadamard product, x_t is the input features of t, h_{t-1} and h_t are the hidden state of $t-1$ and t time step, respectively, f, i, $\tilde{C}$ and O represent the forget gate, the input gate, the cell candidate state, and the output gate, respectively, W and R are the input weight matrix and the recurrent weight matrix, respectively, and b is the bias vector.

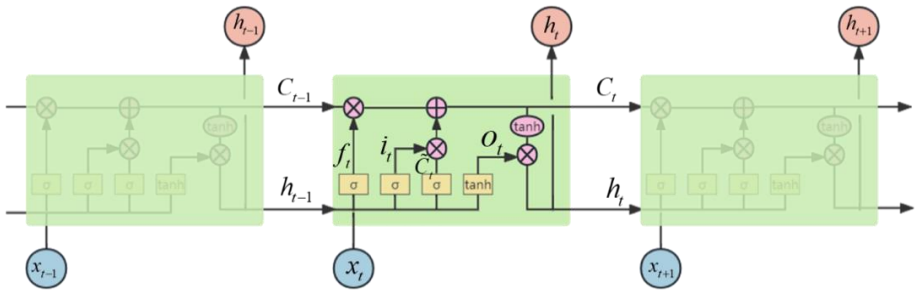

Figure S1. The structure of LSTM

2.2 GRU

GRU is a neural network model proposed by Chung et al., (2014) [8]. Like LSTM, GRU operates by utilizing special gate structures to facilitate information transfer. However, the structure of GRU is simpler than LSTM, and it only has two gates (i.e., Update Gate and Reset Gate) (as shown in Figure S2). At the time step t, the GRU has two input variables: the input features x_t and hidden state h_{t-1}, and one output variable: the hidden state h_t. The mathematical expressions of the GRU are as follows:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \tag{S4}$$

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \tag{S5}$$

$$\tilde{h}_t = \tanh(W_{xh}x_t + W_{hh}(h_{t-1} \odot r_t) + b_h) \tag{S6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{S7}$$

where σ and tanh are the activation functions, $\tilde{h}_t$ are the hidden candidate state, W and b are the weights and bias, respectively.

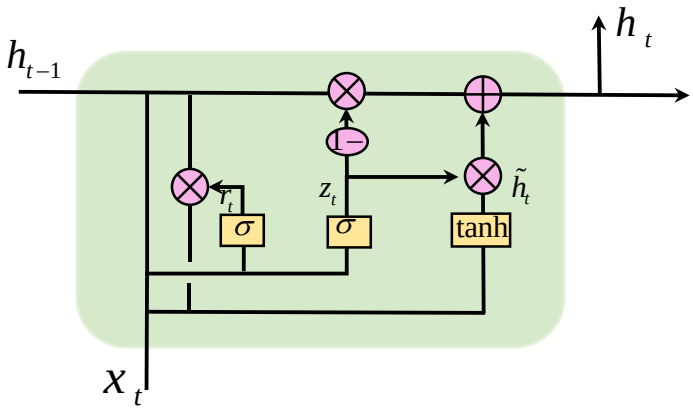

Figure S2. The structure of GRU

2.3 MLP

MLP is a typical feed-forward Artificial Neural Network (ANN), mainly composed of three parts: input layer, hidden layer, and output layer. The input layer represents the input features, while the hidden layer can be a single layer or multiple layers. When MLP has multiple hidden layers, the output of the previously hidden layer becomes the input of the next hidden layer. Depending on the application, the output layer can be a sigmoid function or a linear function [9–10]. MLP can be used for classification, regression, and simply unsupervised learning, and is suitable for classification of large amounts of data and establishing complex nonlinear mappings, etc. MLP can be simply described as a mapping from input features x to output features y [11], with the mathematical formula as follows [12]:

$$y = f(x) + \varepsilon \tag{S8}$$

where f represents the mapping and ε is the process noise.

The hidden layer consists of K hidden neurons, and each neuron is a weighted sum of predictors [11]:

$$a_k = \sum_{i=1}^{M} w_{ki}^{(1)} x_i + w_{k0}^{(1)}, k = 1, \ldots, K \tag{S9}$$

where a_k is a hidden neuron, $\{w_{ki}^{(1)}\}_{i=1}^{M}$ are unknown weights associated with each input neuron, $w_{k0}^{(1)}$ is an unknown bias used to correct the estimation bias; the superscript in Eq. (S9). indicates the number of the hidden layer. The rule of thumb is that the number of hidden neurons should be half the number of predictors and should never be more than twice as large [13]. The output of the hidden neuron is calculated by passing Eq. (S9) to the transfer function:

$$z_k = \psi\left(a_k\right), k = 1, \ldots, K \tag{S10}$$

Where z_k is the output and ψ is the logistic sigmoid transfer function, ranging from 0 to 1. Finally, the connection between the hidden layer and the output layer is established using a linear transfer function.

$$y_j = \sum_{k=1}^{K} w_{jk}^{(2)} z_k + w_{j0}^{(2)} \tag{S11}$$

Where y_j is the output neuron, i.e., the predicted value of the model (j =1, ……, J); $\{w_{jk}^{(2)}\}_{k}^{K}$ and $w_{j0}^{(2)}$ represent the unknown weights and bias of the output layer, respectively. During the training period, the MLP model solves the unknown parameters of Eqs. (S9) and (S10) by backpropagation [9].

3. Three Methods for Integrating GRACE/GRACE-FO and Wells Data to Downsclae GWSA

3.1 Method 1

The downscaling steps for Method 1 are as follows:

1) **The features and targets:** The downscaled feature variables are ERA5-Land P, GLAEM ET, CLSM SMSA, CLSM SNSA, CLSM CNSA, and CLSM temperature. All the feature variables have two spatial resolutions: 0.5°×0.5° and 0.25°×0.25° The target variables are the 6 GRACE-derived GWSAs. And the target variables only one spatial resolution, which is 0.5°×0.5°.

2) **Splitting the training and testing datasets:** The training set consists of time series data from 21 grid cells (~70%), and the test set comprises time series data from the remaining 9 grid cells (~30%) (Figure S3). Both training and testing periods span from January 2004 to December 2021.

3) **DL model training and testing:** The training dataset comprises 0.5°×0.5° features and 0.5°×0.5° targets for the 21 grid cells mentioned above, covering the time range from January 2004 to December 2021. The testing datasets includes 0.5°×0.5° features and 0.5°×0.5° targets for the 9 grid cells mentioned above. The LSTM model using the training dataset, validated with the testing dataset to ensure proper training, and the well-trained model is saved for further use.

4) **Predicting the 0.25 target:** Inputting the 0.25°×0.25° features of all grid cells into the well-trained model to predict 0.25°×0.25° GRACE-derived GWSA $GWSA_{DownscaleSIM}$.

5) **Error estimation:** The errors from both the testing (Eq (S12)) and the training (Eq (S13)) are concatenated by latitude and longitude, forming a complete error dataset for the 30 grid cells. Then, using Kriging interpolation, the 0.5°×0.5° errors are interpolated to 0.25°×0.25° resolution (Eq (S14)) (Figure S3), serving as the error for the downscaling model $error_{Downscale}$. The downscaling results are obtained by adding the model-predicted 0.25°×0.25° GRCAE-derived GWSA ($GWSA_{DownscaleSIM}$) to the estimated model error ($error_{Downscale}$) (Eq (S14)). Note: Since there are six target variables, the step 2) to 5) is repeated six times to obtain six different GRACE-derived GWSAs.

$$error_{Trian} = GWSA_{TrainTruth} - GWSA_{TrainSIM} \tag{S12}$$

$$error_{Test} = GWSA_{TestTruth} - GWSA_{TestSIM} \tag{S13}$$

$$error_{Downscale} = \begin{cases} Kriging(error_{Train}), & the\ grid\ cell\ in\ training\ set \\ Kriging(error_{Test}), & the\ grid\ cell\ in\ testing\ set \end{cases} \tag{S14}$$

$$GWSA_{Downscale} = \begin{cases} GWSA_{DownscaleSIM} + error_{Downscale}, & Method\ 1,2 \\ GWSA_{DownscaleSIM}, & Method\ 3 \end{cases} \tag{S15}$$

Where, $GWSA_{TrainTruth}$ represents the true values of the training set's target, $GWSA_{TrainSIM}$ represents the simulated values of the training set's target variable. $GWSA_{TestTruth}$ represents the true values of the testing set's target variable, and $GWSA_{TestSIM}$ represents the simulated values of the testing set's target variable. $error_{Downscale}$ represents the estimated downscaling model error. $GWSA_{DownscaleSIM}$ is the 0.25°×0.25° GRACE-derived GWSA simulated by the deep learning model.

6) **Selecting the best downscaling result:** At the regional average scale, calculate the correlation coefficients between the six downscaled GRACE-derived GWSAs and the monthly in-situ GWSA for Beijing (spanning from January 2007 to December 2021). Identify the two datasets with the highest correlation, which are JPL Mascon and GSFC Mascon. Compute the average of the 0.25°×0.25° JPL Mascon GWSA and GSFC Mascon GWSA as the final downscaled result.

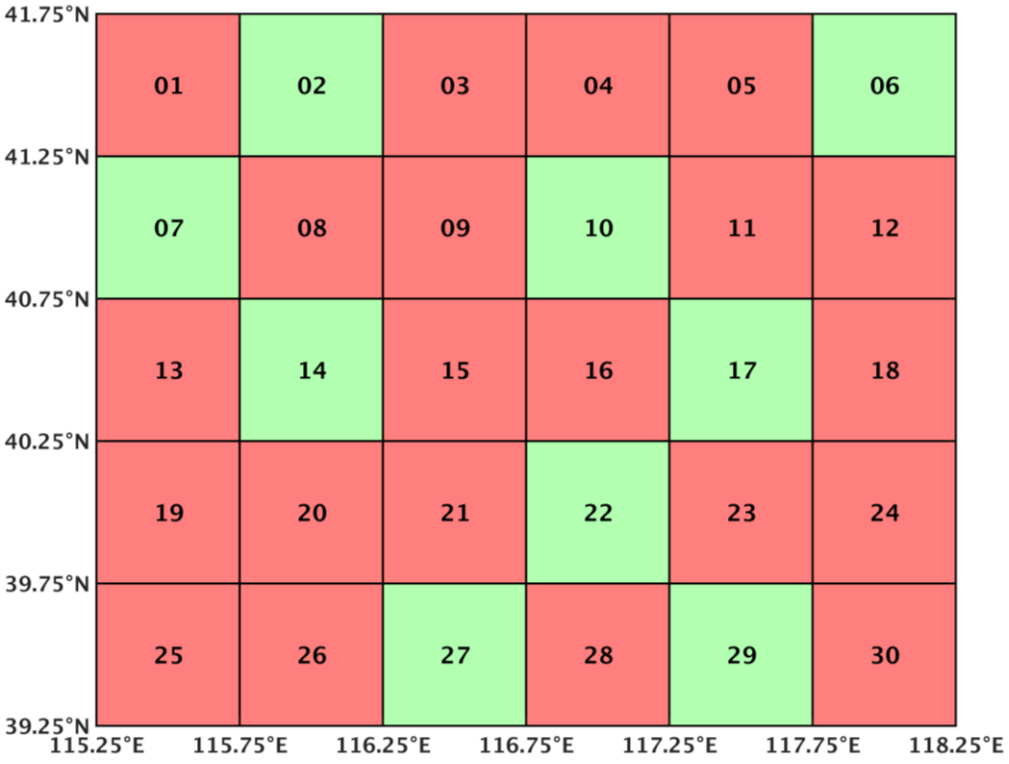

Figure S3. The red grids represent the training grids, the green represents the testing grids, and the number on each grid indicates the grid index.

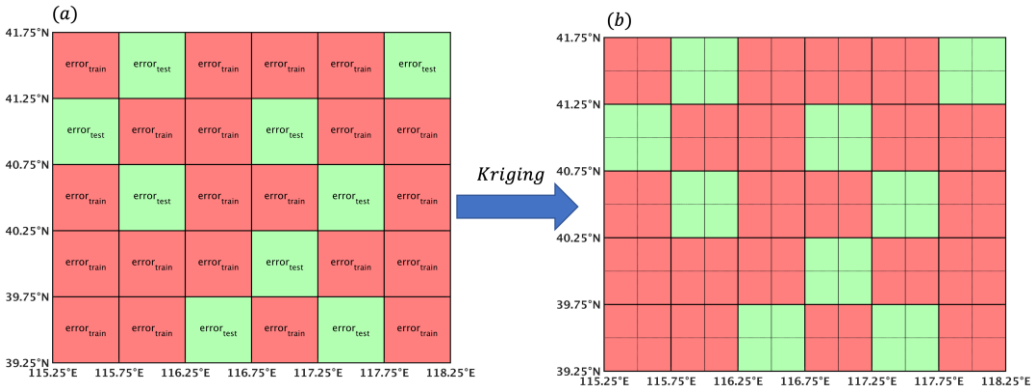

Figure S4. (a) The $error_{Train}$ and $error_{Test}$ at the 0.5°×0.5° spatial resolution. The red grids represent the $error_{Train}$, the green represents the $error_{Test}$. (b) The estimated model error of downscaled $error_{Downscale}$ at the 0.25°×0.25° spatial resolution. Kriging represents Kriging interpolation.

3.2 Method 2

The downscaling steps for Method 2 are as follows:

1) **Making the target variable:** Extract 6 different 0.5°×0.5° GRACE-derived GWSAs (CSR Mascon, GSFC Mascon, JPL mascon, CSR SH, FZ SH, JPL SH) the grids of 14 districts in Beijing and calculate the average value of these 6 datasets in each district. At the regional average scale, compute the correlation of each 0.5°×0.5° GRACE-derived GWSA with the in-situ GWSA for that district. Choose the 0.5°×0.5° GRACE-derived GWSA with the highest correlation as the district's GRACE-derived GWSA and record the score of that dataset. Calculate the data score as follows: for example, in Changping District, if JPL Mascon GWSA has the highest correlation with the in-situ GWSA, it gets a score of one. If JPL Mascon GWSA still has the highest correlation in Chaoyang District, it gets another point. For grids outside the 14 districts, select the dataset with

the highest score among the 6 GRACE-derived GWSAs as the GRACE-derived GWSA. Finally, concatenate the selected GWSAs based on their coordinates to create the complete 0.5°×0.5° GRACE-derived GWSA, denoted as GRACE-derived $GWSA_{concatenate}$.

The 14 counties are Changping, Chaoyang, Daxing, Fangshan, Fengtai, Haidian, Huairou, Mentougou, Miyun, Pinggu, Shijingshan, Shunyi, Tongzhou, and Yanqing.

2) **The features and targets:** The downscaled feature variables are ERA5-Land P, GLAEM ET, CLSM SMSA, CLSM SNSA, CLSM CNSA, and CLSM temperature. All the feature variables have two spatial resolutions: 0.5°×0.5° and 0.25°×0.25°. The target variable is GRACE-derived $GWSA_{concatenate}$ in step 1), and its spatial resolution is 0.5°×0.5°.

3) **Splitting the training and testing datasets:** The training set consists of time series data from 21 grid cells (~70%), and the test set comprises time series data from the remaining 9 grid cells (~30%) (Figure S2). Both training and testing periods span from January 2004 to December 2021.

4) **DL model training and testing:** The training dataset comprises 0.5°×0.5° features and 0.5°×0.5° targets for the 21 grid cells mentioned above, covering the time range from January 2004 to December 2021. The testing datasets includes 0.5°×0.5° features and 0.5°×0.5° targets for the 9 grid cells mentioned above. The LSTM model using the training dataset, validated with the testing dataset to ensure proper training, and the well-trained model is saved for further use.

5) **Predicting the 0.25 target:** Inputting the 0.25°×0.25° features of all grid cells into the well-trained model to predict 0.25°×0.25° GRACE-derived GWSA $GWSA_{DownscaleSIM}$.

6) **Error estimation:** The errors from both the testing (Eq (S12)) and the training (Eq (S13)) are concatenated by latitude and longitude, forming a complete error dataset for the 30 grid cells. Then, using Kriging interpolation, the 0.5°×0.5° errors are interpolated to 0.25°×0.25° resolution (Eq (S14)) (Figure S4), serving as the error for the downscaling model $error_{Downscale}$. The downscaling results are obtained by adding the model-predicted 0.25°×0.25° GRCAE-derived GWSA ($GWSA_{DownscaleSIM}$) to the estimated model error ($error_{Downscale}$) (Eq (S15)).

3.3 Method 3

The downscaling steps for Method 3 are as follows:

1) **Making the features variable:** At the regional average scale, the correlations between the six 0.5°×0.5° GRACE-derived GWSAs and in-situ GWSA for Beijing were calculated. The one with the highest correlation was selected for downscaling, which turned out to be the JPL Mascon GWSA. The 0.5°×0.5° JPL Mascon GWSA was linearly interpolated to 0.25°×0.25°, denoted as JPL Mascon GWSA'. The 0.5°×0.5° JPL Mascon GWSA and JPL Mascon GWSA' are among the features used in the DL model.

2) **Making the target variable:** The in-situ groundwater levels (GWL) from the 41 wells are transformed into in-situ groundwater storage anomaly (GWSA). For each well, the process involves subtracting its mean value ($\overline{GWL}$) over the entire time period,

resulting in GWLA (Eq (S16)). This value is then multiplied by the specific yield (Sy) to obtain the in-situ GWSA for that well (Eq (S17)). The time range for the in-situ GWSAs from the 41 wells spans from January 2006 to December 2016. These in-situ GWSAs from the 41 wells serve as the target data.

$$GWLA(t) = GWL(t) - \overline{GWL} \tag{S16}$$

$$GWSA = GWLA * Sy \tag{S17}$$

Where, GWLA is the groundwater level anomaly for each well, $\overline{GWL}$ is the averaged value for the entire time period, and Sy is specific yield, (Sy=0.06).

3) **The features and targets:** The downscaled feature variables are ERA5-Land P, GLAEM ET, CLSM SMSA, CLSM SNSA, CLSM CNSA, CLSM temperature, JPL Mascon GWSA, and JPL Mascon GWSA'. Except for JPL Mascon GWSA and JPL Mascon GWSA', the other feature variables have two spatial resolutions: 0.5°×0.5° and 0.25°×0.25°. The spatial resolution of JPL Mascon GWSA is 0.5°×0.5°, and the JPL Mascon GWSA' is 0.25°×0.25°. The target variable is the in-situ GWSAs from 41 groundwater wells.

4) **Splitting the training and testing datasets:** The training set consists of time series data from 30 wells (~73%), and the test set comprises time series data from the remaining 11 wells (~27%). Both training and testing periods span from January 2006 to December 2016.

5) **DL model training and testing:** The training dataset comprises 0.5°×0.5° features and in-situ GWSAs from 30 wells mentioned above, covering the time range from January

2006 to December 2016. The testing datasets includes 0.5°×0.5° features and in-situ GWSAs from 9 wells mentioned above. The LSTM model using the training dataset, validated with the testing dataset to ensure proper training, and the well-trained model is saved for further use.

6) **Predicting the 0.25 target:** Inputting the 0.25°×0.25° features of all grid cells into the well-trained model to predict 0.25°×0.25° GRACE-derived GWSA ($GWSA_{DownscaleSIM}$). $GWSA_{DownscaleSIM}$ represents the final downscaled outcomes (Eq (15)).

Note: The 41 groundwater wells are concentrated in the southeastern part and cannot be interpolated into a complete 0.25° grid, so error estimation is not performed for Method 3.

4. Accuracy Evaluation Metrics

In this study, three commonly used metrics were applied to evaluate the performance of the model: the Pierre correlation coefficient (CC), the Nash-Sutcliffe efficiency coefficient (NSE), and root mean square error (RMSE).

$$CC = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(\hat{y}_i - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}} \quad \text{(S18)}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{S19}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{S20}$$

where n is the number of months of training or testing or reconstruction, y_i and $\hat{y}_i$ represent the observed value and model simulation at month i, respectively. $\overline{y}$ and $\overline{\hat{y}}$ represent the mean of the observed value and simulation value, respectively. The range of CC is -1 to 1, the larger the absolute value of CC, the stronger the correlation, and the sign indicates a positive or negative correlation. NSE is from negative infinity to 1, and relative proximity to 1 indicates a model with accurate simulative skill. In contrast, RMSE closer to 0 suggests a model with better performance.

5. Supplementary Figures

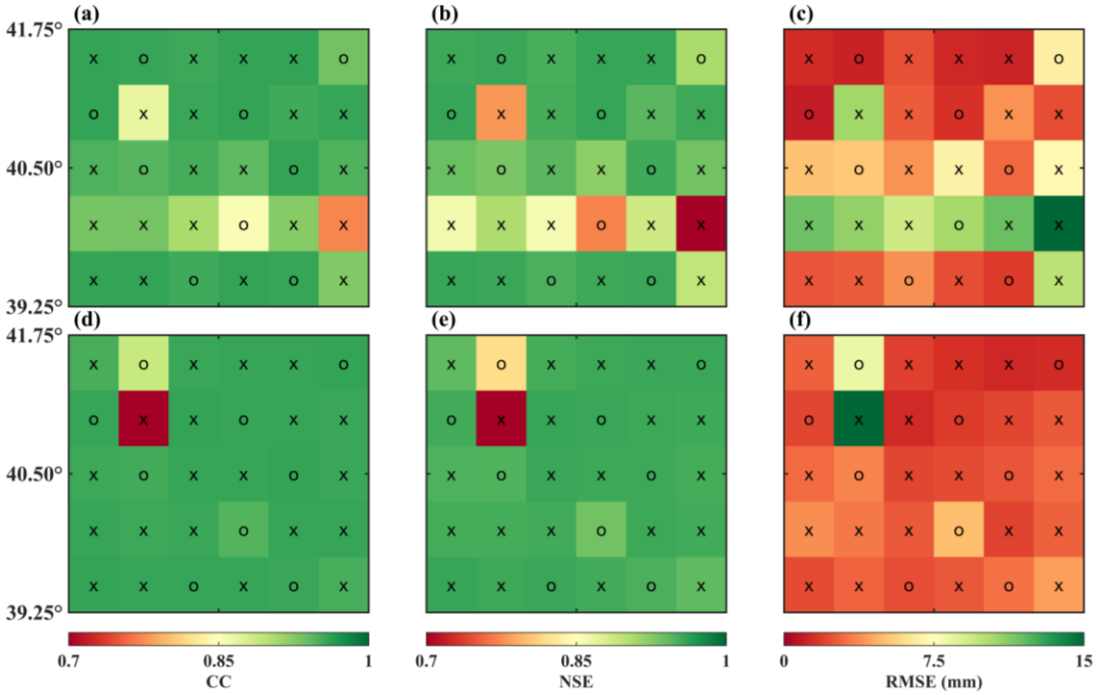

Figure S5. The accuracy of training and testing sets for Method 1 and Method 2. "x" reprents the training grid cells, and "o" represents testing grid cells. (a)~(c) Method 1, (d)~(f) Method 2.

Figure S5 displays the CC, NSE, and RMSE for the training and test sets during the downscaling process. Among these, 21 grid cells (x) represent the accuracy of the training set, while the remaining 9 grid cells (o) represent the accuracy of the test set. For both Method 1 and Method 2, whether in the training or test set, CC and NSE consistently exceed 0.7, and RMSE remains below 15. This indicates that the models are reasonably reliable.

6. Supplementary Tables

Table S1. The accuracy of training set for Method 3

Training Set (Id)	CC	NSE	RMSE (mm)
1	0.97	0.93	33.34
2	0.81	0.56	43.73
3	0.74	-0.70	117.41
4	0.89	0.45	30.45
5	0.99	0.98	19.40
6	0.98	0.95	34.87
7	0.99	0.98	11.19
8	0.98	0.94	17.26
9	0.98	0.81	111.48
10	0.80	0.56	56.00
11	0.99	0.98	25.81
12	0.97	0.90	26.19
13	0.96	0.90	76.71
14	0.94	0.89	64.55
15	0.33	-0.5	143.35
16	0.99	0.98	17.68
17	0.20	-0.6	375.75
18	0.86	0.63	32.15
19	0.99	0.98	58.61
20	0.76	0.50	326.00
21	0.99	0.98	22.21
22	0.99	0.98	21.19
23	-0.10	-2.11	163.59
24	0.97	0.94	13.27
25	0.85	0.60	84.87
26	0.97	0.94	39.39
27	0.75	-1.01	98.12
28	0.20	-1.23	94.21
29	0.90	0.30	211.78
30	0.97	0.89	20.68

Table S2. The accuracy of testing set for Method 3

Testing Set (Id)	CC	NSE	RMSE (mm)
1	0.85	0.60	148.12
2	0.91	0.45	81.45
3	0.99	0.99	17.30
4	0.70	0.40	192.75
5	0.90	0.70	137.11
6	0.79	-1.20	171.48
7	-0.79	0.72	41.62
8	0.90	0.61	62.95
9	0.98	0.97	62.81
10	0.15	0.02	280.45
11	0.97	0.92	36.44

Table S1 and Table S2 present the CC, NSE, and RMSE for the training and test sets during the downscaling process of Method 3. For Method 3, we randomly selected 30 of the 41 wells' in-situ GWSAs for the training set, with the remaining 11 wells' in-situ GWSAs designated as the test set. Table S1 displays the CC, NSE, and RMSE between the in-situ GWSAs of the 30 wells in the training set and the simulated in-situ GWSAs from the downscaling model. Overall, the model shows a good fit in the training set. Only a few well locations exhibit less satisfactory fits, for instance, Id: 27 and 28, where NSE exceeds -1. On the testing set (Table S2), the accuracy remains relatively high, with most CC values exceeding 0.70 and most NSE values surpassing 0.60.

References

1. Swenson, S.; Wahr, J.; Milly, P.C.D. Estimated Accuracies of Regional Water Storage Variations Inferred from the Gravity Recovery and Climate Experiment (GRACE). *Water Resour. Res.* **2003**, *39*, 1223. https://doi.org/10.1029/2002WR001808.
2. Cheng, M.; Tapley, B.D.; Ries, J.C. Deceleration in the Earth's Oblateness. *J. Geophys. Res. Solid Earth* **2013**, *118*, 740–747. https://doi.org/10.1002/jgrb.50058.
3. Swenson, S.; Wahr, J. Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.* **2006**, *33*. https://doi.org/10.1029/2005GL025285.
4. Paulson, A.; Zhong, S.; Wahr, J. Limitations on the Inversion for Mantle Viscosity from Postglacial Rebound. *Geophys. J. Int.* **2007**, *168*, 1195–1209. https://doi.org/10.1111/j.1365-246X.2006.03222.x.
5. Feng, W. GRAMAT: A Comprehensive Matlab Toolbox for Estimating Global Mass Variations from GRACE Satellite Data. *Earth Sci. Inform.* **2019**, *12*, 389–404. https://doi.org/10.1007/s12145-018-0368-0.
6. Long, D.; Pan, Y.; Zhou, J.; Chen, Y.; Hou, X.; Hong, Y.; Scanlon, B.R.; Longuevergne, L. Global Analysis of Spatiotemporal Variability in Merged Total Water Storage Changes Using Multiple GRACE Products and Global Hydrological Models. *Remote Sens. Environ.* **2017**, *192*, 198–216. https://doi.org/10.1016/j.rse.2017.02.011.
7. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
8. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
9. Long, D.; Shen, Y.; Sun, A.; Hong, Y.; Longuevergne, L.; Yang, Y.; Li, B.; Chen, L. Drought and Flood Monitoring for a Large Karst Plateau in Southwest China Using Extended GRACE Data. Remote Sens. Environ. 2014, 155, 145–160. https://doi.org/10.1016/j.rse.2014.08.006.
10. Pal, S.K.; Mitra, S. Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Trans. Neural Netw.* **1992**, *3*, 683–697. https://doi.org/10.1109/72.159058.
11. Bishop, C. *Pattern Recognition and Machine Learning (Information Science and Statistics*; Springer: New York, NY, USA, 2007.
12. Kumar, K.S.; AnandRaj, P.; Sreelatha, K.; Sridhar, V. Reconstruction of GRACE terrestrial water storage anomalies using Multi-Layer Perceptrons for South Indian River basins. *Sci. Total Environ.* **2023**, *857*, 159289, doi: 10.1016/j.scitotenv.2022.159289.
13. Berry, M.J.A.; Linoff, G.S. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2009.