Toward improved lumped groundwater level predictions at catchment scale: mutual integration of water balance mechanism and deep learning method

Hejiang Cai[1,2,3], Suning Liu[4], Haiyun Shi[2,3,*], Zhaoqiang Zhou[2,3], Shijie Jiang[5], Vladan Babovic[1]

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

[2]State Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China

[3]Guangdong Provincial Key Laboratory of Soil and Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China

[4]Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea

[5]Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research, Leipzig, Germany

Corresponding author: Dr. Haiyun Shi (shihy@sustech.edu.cn)

Abstract

Model development in groundwater simulation and physics informed deep learning (DL) has been advancing separately with limited integration. This study develops a general hybrid model for groundwater level (GWL) simulations, wherein water balance-based groundwater processes are embedded as physics constrained recurrent neural layers into prevalent DL architectures. Because of the automatic parameterizing process, physics-informed deep learning algorithm (DLA) equips the hybrid model with enhanced abilities of inferring geological structures of catchment and unobserved groundwater-related processes implicitly. The main purposes of this study are: 1) to explore an optimized data-driven method as alternative to complicated groundwater models; 2) to improve the awareness of hydrological knowledge of DL model for lumped GWL simulation; and 3) to explore the lumped data-driven groundwater models for cross-region applications. The 91 illustrative cases of GWL modeling across the middle eastern continental United States (CONUS) demonstrate that the hybrid model outperforms the pure DL models in terms of prediction accuracy, generality, and robustness. More specifically, the hybrid model outperforms the pure DL models in 78% of catchments with the improved ΔNSE=0.129. Meanwhile, the hybrid model simulates more stably with different input strategies. This study reveals the superiority and powerful simulation ability of the DL model with physical constraints, which increases trust in data-driven approaches on groundwater modellings.

Key words

Groundwater level predictions; Water balance mechanism; Deep learning; Catchment scale

Highlights

- A novel hybrid model for simulating groundwater level was developed
- The hybrid model integrated water balance equations with deep learning algorithm
- The proposed model presented the superiority and powerful simulation ability
- The automatic parameterizing ability enhanced the model for cross-region simulation

1 Introduction

Groundwater plays an important role in geophysical and hydrological simulation (Cuthbert et al., 2019; Li et al., 2017; Zipper et al., 2018). Two major focuses related to groundwater simulation are: 1) exploring the essence of hydrological cycle and its potential influences on geophysical dynamic system, such as climate change (Ma et al., 2021a; Fan et al., 2013; Mohan et al., 2018) and drought propagation (Ma et al., 2021b; Sadeghfam et al., 2018; Zhou et al., 2020); and 2) exploring rational strategies including management, utilization and protection of groundwater resources under the impacts of human activities, such as pumping (Zipper et al., 2018) and pollutant transport (Ossai et al., 2020; Tian et al., 2015). For groundwater-related simulations, seeking accurate prediction of groundwater level (GWL) is an inevitable task. Due to the deepening understanding about the essence of geophysical dynamic system, many efforts have been devoted into process-based hydrological models (Feng et al., 2020). In general, the successes of traditional groundwater-related hydrological models relied heavily on detailed geological attributes and meteorological data. More specifically, referring to collected local hydrogeological information, such as boundary conditions, land covers, and so on, a catchment is divided into several hydrological units and water interaction between adjacent hydrological units is simulated numerically (Mohan et al., 2018). Furthermore, strong mathematical derivation ability is required for establishing groundwater-related models since the relevant governing equations are almost two- or three-dimensional partial differential equations (PDEs). In general, process-based models are the most suitable simulation strategies for catchments where the knowledge of flow path, geological characteristics and boundary conditions are vital and extensively clear (Sahoo et al., 2017). Despite there are many successful cases of applying traditional process-based models to

simulate GWL in specific regions, it has been criticized as too complex, parametric, difficult and regional limited to use (Beven and Cloke, 2012; Clark et al., 2015a; Simone Fatichi et al., 2016; Tran et al., 2021). Moreover, it is still extremely difficult for process-based models to perfectly describe the water flow with nonlinear interactions, spatial heterogeneity, and time lags in real groundwater systems (Clark et al., 2015b). With the continuous development of computer science and deep learning algorithms (DLAs), data-driven hydrological models seem to be reliable alternatives for traditional process-based hydrological models.

In general, the main data sources for hydrological modeling originate from observations, surveys, and lab experiments over years of systematic research (Chadalawada et al., 2020). Multiple variables, sources, and resolutions in space and time doom the complex and big-data nature of hydrological data. With the aiming of exploring, extracting, and simulating the information from the raw data, different types of DLAs emerged as times required. Among the DL strategies, artificial neural networks (ANNs) and their optimized forms have become the mainstream methods for establishing data-driven hydrological models. During the past decades, applications of data-driven DL models have burgeoned in different fields of hydrology and their performances have been widely recognized (Reichstein et al., 2019). For examples, commonly recognized DLAs for hydrological timeseries modeling, such as streamflow and GWL forecasting, include: recurrent neural network (RNN) based models, for instance, long short-term memory (LSTM) (Feng et al., 2020; Gao et al., 2020; Kratzert et al., 2019; Jiang et al., 2021) and gate recurrent unit (GRU) (Cai et al., 2021; Zhang et al., 2021); convolutional neural networks (CNN) (Jiang et al., 2018; Jiang et al., 2019); genetic programming (GP) (Babovic, 2009; Babovic and Keijzer, 2002; Chadalawada et al., 2017; Liu and Shi, 2019; Chadalawada et al., 2020); and etc. Admittedly, benefited from the outstanding computational efficiency with diverse algorithm types,

DL models present excellent adaptability and versatility when serving as alternatives to traditional models. Comparing to the process-based models, DL models have advantages in terms of easy construction, computing speed, and data requirements (Reichstein et al., 2019). However, criticizes on DL models focus on the black-box essence, which means that DL models are more like computing tests without any physical meanings. As a result, DL models have not been psychologically accepted by hydrologists because such models cannot improve our understanding the essence of natural hydrogeological processes at present.

In recent years, application of artificial intelligence methods on hydrological modeling has developed into a new and critical stage which integrating DLAs with dynamic geophysical processes is expected to enhance the performance and generality of data-driven models simultaneously (Frame et al., 2022; Jiang et al., 2020; Hoedt et al., 2021; Reichstein et al., 2019; Zhao et al., 2019). Recent studies have demonstrated two strategies of integrating physical constraints and DLAs: 1) using interpretation methods to demonstrate the prevalent DL architectures for gaining scientific insights. For example, Jiang et al. (2022) explored the flood inducing factors by analyzing inner works of the LSTM models. 2) embedding physical processes into DL models to improve their awareness to systematic dynamic processes. Among these research works, one strategy is adding physical informed equations into loss functions of deep learning models (Raissi et al., 2019). For example, Wang et al. (2020) proposed a Theory-guided Neural Network (TgNN), which considered groundwater related factors (including governing equations, boundary conditions, initial conditions, engineering controls and expert experience) as residual terms for loss function of the neural networks. This type of strategy is mostly suitable for specific spatial problems as it could be considered as a powerful approximator for the PDEs. Another strategy is adding physical constraints into the inner neural networks for forward

propagation. This type of strategy is majorly based on recurrent neural networks (RNN) for timeseries simulations. For example, Kratzert et al. (2019) optimized the LSTM models with regional entity awareness for streamflow simulation; Zhao et al. (2019) integrated Penman-Monteith equations into ANN models for simulation of evapotranspiration. Significantly, Niu et al. (2019) demonstrated the relationship between the network architecture of RNN family and numerical method, and theoretically supported the use of RNN to solve problems involving system dynamics. Jiang et al. (2020) successfully embedded snowmelt process into DL models for streamflow simulations. In general, most previous studies only focused on streamflow-related tasks, applications of physically constrained DL models on the GWL simulation remains to be explored.

With the purpose of improving the performance of lumped DL model for GWL simulation at catchment scale with limited geological observations, this study proposed a hybrid hydrological model with water balance as physical constraints and DLA as cornerstone. The embedding of water balance equations is theoretically supported by the algorithm of solving ordinary differential equations (ODEs) with RNN (Jiang et al., 2020). The groundwater-related water balance equations are summarized from GSFLOW model (Markstrom et al., 2008), EXP-HYDRO model (Patil & Stieglitz, 2014), and TOPMODEL (Kirkby, 1975). Two prevalent DLAs for timeseries simulations, i.e., one dimensional convolutional neural network (1D-CNN) and gate recurrent unit (GRU), are established for comparisons of simulating performance. Specifically, the hybrid model consists of a self-designed RNN model (WB-Model) with wrapped water balance equations and a prevalent DL model (two-layer 1D-CNN model). Self-designed parameters of water balance equations with physical meanings in WB-Model are determined during the iterative (training) process of DLA, which strengthens the model's understanding of physical process and

hydrogeological characteristics of target catchments. For hybrid model, GWL simulated form WB-Model serves as an input for a 2-layer 1D-CNN model to give the final output. As a result, enhanced simulation accuracy, robust, generalization ability and intelligence for inferring characteristics of catchments are expected from the proposed hybrid model. Overall, this study demonstrates that the novel hybrid model can garner the GWL-related physical knowledge in a catchment vision if integrated with physical constraints properly, which makes the physical constrained DL model be more accurate, interpretable, feasible, advanced, and promising in terms of GWL simulations for cross-region and less ungauged catchments.

2 Data and methods

2.1 Study area and multisource data

Continental United States (CONUS), as shown in subplot(a) of Fig. 1, is divided into 18 major watersheds by U.S. Geological Survey (https://www.usgs.gov/), and each watershed contains either drainage area of a major river or the combined drainage area of several rivers. Experimental data collected in this study consist of two datasets from 10 major watersheds: New England Region (01), Mid Atlantic Region (02), South Atlantic-Gulf Region (03), Great Lakes Region (04), Ohio Region (05), Tennessee Region (06), Upper Mississippi Region (07), Lower Mississippi Region (08), Missouri Region (10) and Arkansas-White-Red Region (11). The first dataset is from publicly available Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS), which contains two types of data to describe a specific catchment with minimal human disturbance: 1) seven types of basin-averaged daily hydrometeorological timeseries data mostly recorded from 1980 to 2014 at hydrometeorological observation stations: precipitation (P), surface downward solar radiation (SRAD), snow water equivalent (SWE), maximum temperature (T_{max}), minimum temperature (T_{min}), near surface daily average vapor pressure (V_p), and streamflow observations

at catchment outlet (SF); and 2) six types of averaged catchment attributes, i.e., topography and location, climate indices, hydrological signatures, land cover characteristics, soil characteristics, and geological characteristics (Addor et al., 2017; Newman et al., 2015). Each catchment in the CAMELS dataset is represented by a hydrological unique code (HUC). The second dataset is the freely available daily GWL data corresponding to catchments in the CAMELS dataset, which was collected and compiled by U.S. Geological Survey. In this study, we selected the GWL data following three principles simultaneously. First, the selected groundwater wells must provide daily GWL data with 25 to 30 consecutive years between 1980 and 2014. Lower limit of 25 years was set to ensure both enough monitoring wells that could meet the standard and a satisfying amount of data for the need of DL model. Meanwhile, upper limit of 30 years was set to control the difference of data volume for DL model to ensure the simulation results were comparable. Second, as shown in Fig. 1, GWL data must be collected from the monitoring wells closest to the streamflow observation stations from the same catchment. This principle was set to make sure that the hydrological factors in the CAMELS dataset were major driving forces of the GWL changes. Despite we recognized that the boundaries of groundwater and streamflow catchments are not always perfectly overlapped, among the dataset we collected, compared with the average catchment area of 812 km^2, we believed that the average distance of 20 km between groundwater monitoring wells and streamflow stations was small enough to ensure their potential interactions. Third, since the CAMELS dataset contained only catchments with minimal human impacts and human impacts were not considered in this study, the groundwater data we collected should also avoid urban and agricultural areas with frequent human activities. However, most of the wells we collected in the western CONUS, especially in California Region, presented obvious traces of human activities such as pumping. The number of eligible wells in those regions is negligible. As

a result, 91 catchments from 10 major watersheds located in the central eastern CONUS were used for this study (Fig. 1).

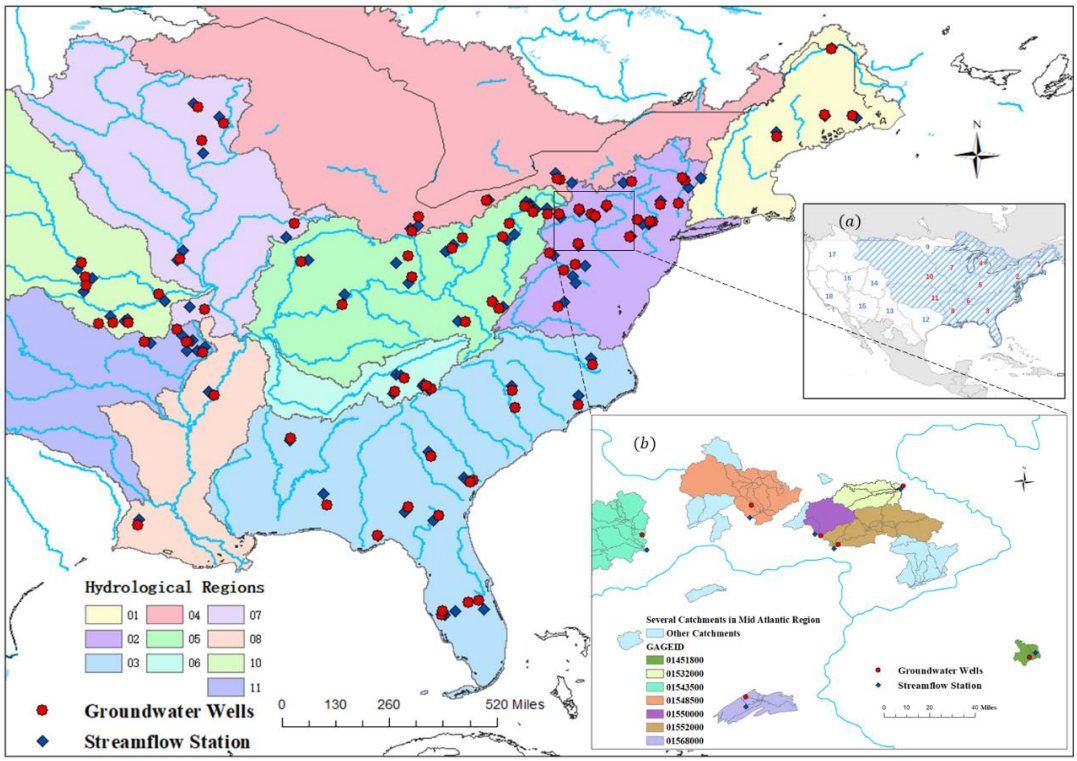

Fig. 1. Distribution of streamflow observations and groundwater monitoring wells used in this study. Map of the 18 major watersheds of Continental United States (subplot(a)). Examples of several catchments and locations of groundwater wells and streamflow stations (subplot(b)).

2.2 Deep learning methods for hydrological timeseries simulation

2.2.1 Gate recurrent unit

A general concept of RNNs refers to a class of ANNs with recurrent cells, which works on the principle of storing the output of a particular layer and feeding it back to the input in order to predict the optimate output of the layer. During the past decades, a variety of RNN-based DL models, including GRU and LSTM, were proposed following different strategies of designing the

recurrent cell. To better introduce GRU model, we will firstly introduce the basic model of RNN family, i.e., simple RNN model. As shown in Fig. 2(a), a simple RNN (Rumelhart et al., 1986) recurses in the evolution direction of the sequence and all nodes are connected in a chain. The key point of a simple RNN model is the concept of hidden state (S_t), which stores information from previous time steps and delivers the previous features to predict the output (y_t). The specific algorithms are listed as follows:

$$S_t = \sigma_s(U_s x_t + W_s S_{t-1} + b_s) \tag{1}$$

$$y_t = \sigma_y\left(W_y S_t + b_y\right) \tag{2}$$

where σ is an activation function strategy, U and W are weight matrices, and b is a bias vector. Theoretically, RNN can make use of all the information from former sequence, which makes it a preferred strategy for timeseries analysis. However, due to the gradient vanishing and explosion problems, a simple RNN would perform poorly when it comes to long sequence analysis because outputs are likely to be only determined by several former steps. In order to solve these problems, LSTM was firstly brought up for language processing with three gating controllers and two hidden states (Hochreiter and Schmidhuber, 1997). To simplify the LSTM structure, GRU neural networks (as shown in Fig. 2(b)) were proposed by Cho et al. (2014). GRU neural networks share a similar chain structure with that of a simple RNN, but the internal operations inside the recurrent cell are optimized for long short-term sequence simulation. The core algorithm of GRU is the two-gate controller: reset gate and update gate. The reset gate determines how to combine the new input information with the previous memory (which information will be stored or deleted), and the update gate defines the amount of previous memory saved to the current state (which data will be output from the current state). Computational workflow of GRU can be summarized as follows:

$$z_t = \sigma(W_z[x_t, h_{t-1}] + b_z) \tag{3}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{4}$$

$$\widehat{h_t} = \tanh(W_h \cdot x_t + U_h \cdot (r_t \otimes h_{t-1}) + b_h) \quad (5)$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \widehat{h_t} \quad (6)$$

where W and b are weight matrices and bias vectors; z_t and r_t are update and rese gates' activation vectors; $\widehat{h_t}$ is potential update vector; σ and tanh are sigmoid and hyperbolic tangent functions. Recently, several studies have shown that GRU presents similar, if not better, simulation performance with LSTM but a better computational efficiency due to a reduction of structural complexity (Ayzel and Heistermann, 2021; Gao et al., 2020). Therefore, we applied GRU networks as one scenario of GWL simulation. As shown in Fig. 2(d), referring to the modelling strategy introduced by Cai et al. (2021), we adopted two GRU layers with a full connection layer and a dropout layer for the catchment-scale GWL simulation.

2.2.2 One dimensional convolutional neural network (1D-CNN)

Convolutional neural network was firstly proposed by LeCun et al. (1989), who integrated the back propagation algorithm and shared weights into convolutional neural layer. Normally, the basic structure of CNN consists of input layer, convolutional layer, pooling layer, full connection layer, and output layer. As shown in Fig. 2(c), 1D-CNN structure applied in this study adopts convolution layer and pooling layer alternately. A convolution operation involves two steps: 1) multiple learnable convolutional kernels (also known as filters) read the former layer, such as input layer, by sliding on sequence matrix; and 2) the output features of the upper layer are convoluted with the convolution kernel, that is, the dot product operation is performed between the input term and the convolution kernel, and then the results are sent to the activation function to obtain the output features. The essence of convolutional layer is similar to RNN-based DLA, which is extracting the most relevant features of input sequence for output sequence predicting. Comparing with a fully connected layer, the 1D-CNN layers do not require the manipulation of lags concatenation and decrease the memory resources allocated accordingly (Jiang et al., 2020).

Comparing with the RNN-based DL model, the algorithm of 1D-CNN is more concise, which may make its simulation performance not be as good as that of RNN-based model, but its calculation speed will be much faster (Jiang et al., 2020). To this point, 1D-CNN layers are chosen as one section of the hybrid model for efficient testing.

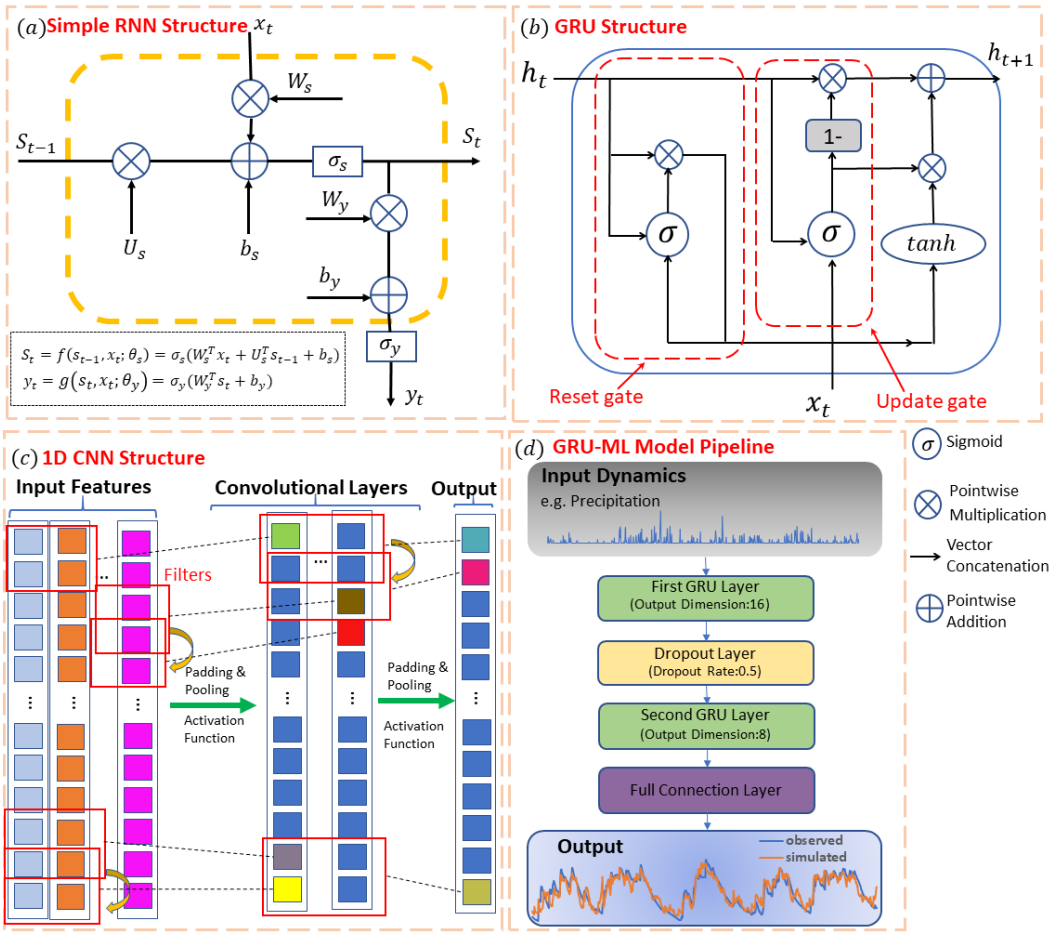

Fig. 2. (a) Structure of a simple Recurrent Neural Network. (b) Structure of Gate Recurrent Unit. (c) Structure of 1D-Convolutional Neural Network. (d) Workflow of GRU-DL model in this study.

2.2.3 DLA for solving ODEs

In general, the simplest form of RNN-based model is that the input vector sequence $(x_1, x_2, \dots, x_t)$ and output vector sequence $(y_1, y_2, \dots, y_t)$ should satisfy the following recursive relationship:

$$y_t = f(y_{t-1}, x_t, t) \tag{7}$$

As shown in Fig. 2, state (S_t) stores and propagates the information from last time step (S_{t-1}), which leads to the simplest RNN model is in fact a recursive computation model as follows:

$$y_t = \tan \mathrm{h}(W_1 y_{t-1} + W_2 x_t + b) \tag{8}$$

where W and b are weight matrices and bias vectors.

Research concerned with dynamics and dynamical systems is usually closely related to the solution of ODEs (Jiang et al., 2020). General ODEs have the following form:

$$\dot{y}(t) = f(y(t), t) \tag{9}$$

Only a few ODEs can be solved analytically. In most cases, numerical method is regarded as the first choice for solving ODEs. Euler method is a basic numerical method for solving initial value problems of the first-order ODEs (Butcher, 2000). Implicit Euler method uses the first-order backward difference quotient instead of differentiation, that is:

$$\dot{y} \approx \frac{y(k) - y(k-1)}{t(k) - t(k-1)} = \frac{y(k) - y(k-1)}{h} \tag{10}$$

where h is time step. In this case, the differential equation becomes an implicit difference equation:

$$\begin{cases} y(k) = y(k-1) + hf\big(t(k), y(k)\big) \\ y(0) = y_0 \end{cases} \tag{11}$$

In the k-th iteration, $y(k-1)$ and $t(k)$ are known values, $y(k)$ is the unknown value to be obtained, and $f\big(t(k), y(k)\big)$ is a function of the unknown value $y(k)$ to be solved. Normally, $y(k)$ can be solved by iterative solution of nonlinear equations (such as Newton's iterative method):

calculating the values of $y(1)$, $y(2)$, $y(3)$, $y(4)$, ..., $y(n)$ in turns, and the discrete sequence is the numerical solution of the differential equation. However, comparing Eq. 7 with Eq. 11, it can be concluded that the Euler solution of ODE is theoretically a special case of RNN. The forward propagation of RNN corresponds to the Euler solution of ODE. Meanwhile, the backpropagation process of RNN for determining the parameters brings out the advantage of applying RNN to solve ODE (Tan et al., 2018). To be more specific, in actual problems, there is a type of problem called "model inference", which is based on the known experimental data. The ultimate goal is to guess the model (mechanism inference) that this batch of data conforms to. The approach to solve this type of problem is roughly divided into two steps: the first step is to guess the form of the model, and the second step is to determine the parameters of the model (Viana et al., 2021). In our case, assuming that this batch of data can be described by an ODE whose form is already known, what needs to be focused on is the estimation of the inner parameters. From the perspective of RNN, the process of parameter determination corresponds to how the propagation mechanism of RNN model is designed: forward propagation is to solve ODE (RNN prediction process), and backpropagation is to infer ODE parameters naturally (RNN training process). An interesting fact is that, inferencing ODE parameters is a well-studied subject, but it is just one of the most basic applications of RNN model.

As shown in Fig. 3, general idea of inserting physical constrained ODEs into the general concept of RNN structure is to replace the step function of simple RNN model with differentiable physical constrained equations. In our case, the proposed WB-Model is designed to encode geosystem dynamics as a neural network architecture with RNN backend. Within the self-designed RNN architecture, the connections (Eq. 1 and Eq. 2) between neurons in simple RNN structure, inputs, states, and outputs, are specified with the state-space representation in an explicit discrete

form of governing functions, which in our case are the water balance equations. Specifically, we replaced the algorithms and the activation functions for calculating the states (S_t) in hidden layer and output (y_t) by water balance ODEs with self-designed state variables, self-designed flux variables and self-learnable parameters (see equations in Fig. 3(a) and Fig. 3(b), respectively). To make it clearer, Fig. 3(c) presents the structure and unfolded network schematic graph of the WB-Model layer. Unlike simple RNN model, which possesses one state variable for each hidden layer, WB-Model is equipped with five states variables ($S_0, S_1, \dots, S_4$) for each of the so-called hidden layer. Each state variable represents a specific groundwater related hydrological reservoir. To calculate and update the state variables, nine flux variables are introduced. Each flux variable represents a flow process. Flux variables are like the 'bridges' connecting the state variables in order. To make the flux variable accurate and meaningful, 16 self-learnable parameters (θ_P), consisting of θ_P^f and θ_P^g, with physical meanings are introduced. Each parameter is determined by the gradient descent algorithm wrapped in the ordinary RNN model. Relevant descriptions about the water balance equations, self-designed state variables, self-designed flux variables, and self-learnable parameters applied in this WB-Model are introduced in detail in subsection 2.3. In summary, we redefined the step function of ordinary RNN to solve the target ODEs by introducing self-designed state variables, self-designed flux variables, and self-learnable parameters with physical meanings.

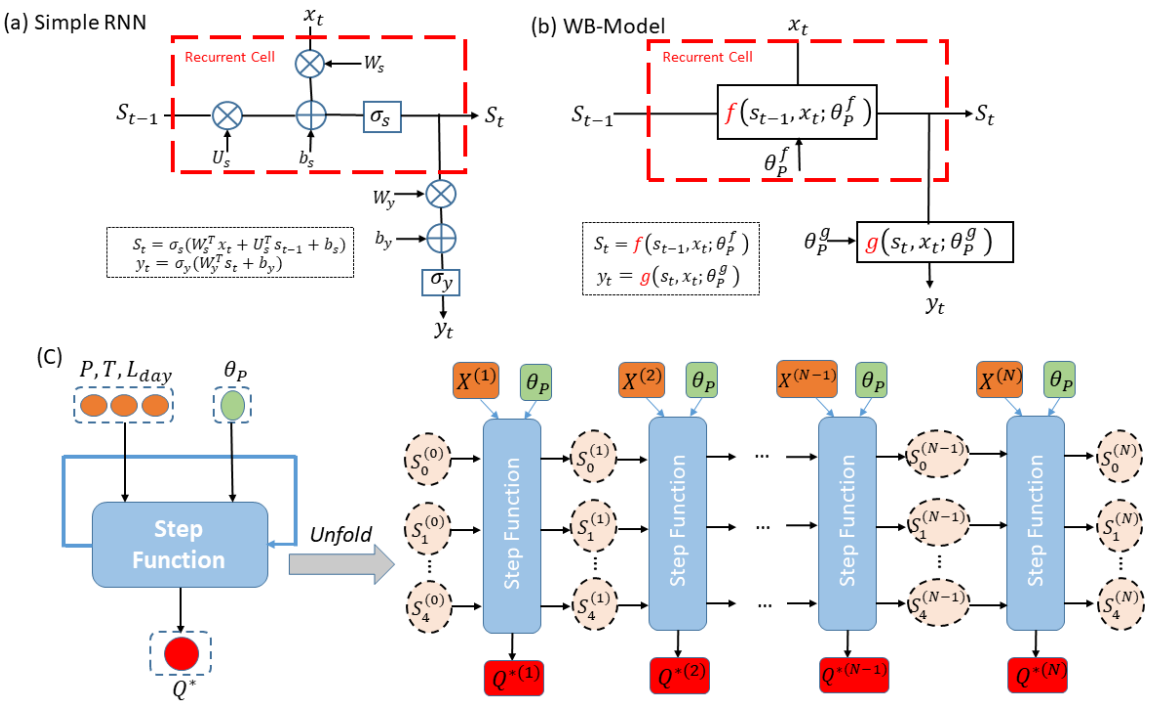

Fig. 3. Architectures of the simple RNN layer and the proposed WB-Model layer.

2.3 Groundwater-related water balance equations

The essence principle for process-based models is water balance, a form of mass conservation law in terms of hydrological research (Herrmann et al., 2016), with a generic equation as:

$$\frac{dS}{dt} = I(t) - O(t) \tag{12}$$

where S is the water storage unit, $I(t)$ is the inflow to the unit at time t and $O(t)$ is the outflow from the unit at time t. Traditional process-based hydrological model usually divides the catchments into segments (hydrological units) and calculates the flow between adjacent segments based on water balance equations and hydrogeological characters of segments. In addition to spatial and temporal resolutions, existing process-based hydrological models majorly differ in mathematical algorithm of $I(t)$ and $O(t)$ (Jiang et al., 2020). In this subsection, we summarized water balance equations related to GWL from different process-based models including GSFLOW model (Markstrom et al., 2008), EXP-HYDRO model (Patil & Stieglitz, 2014) and TOPMODEL (Kirkby,

1975). These processes were integrated into DL by replacing the states and learnable parameters with water balance equations for RNN model. We named the model and equations wrapped in the RNN model as WB-Model, and related processes and equations are listed below.

2.3.1 Canopy interception

As shown in Fig. 4, interception of precipitation by plant canopy is computed during a time step as a function of plant-cover density and the storage available on the predominant plant cover type. The variability of the CAMELS dataset makes it possible to demonstrate the plant canopy interception process since the landcover related data is included (Dunkerley and Booth, 1999). Precipitation that reaches the ground is calculated by the sum of throughfall and precipitation on the catchment not covered by plants. Canopy interception is closely related to the vegetation type and local climate condition (Tao et al., 2020; Trinh and Chui, 2013). In this subsection, we set a reservoir (S_0) for the storage change of canopy interception. Two critical learnable parameters, the maximum capacity of canopy interception (SC_{max}) and canopy interception coefficient related to plant type (K_c), were introduced for physical RNN model. Since CONUS locates in the Northern Hemisphere, we determined the seasons according to the averaged day-length: if the day-length is longer than 0.5, we set a range with higher mean value of K_c and SC_{max} because vegetation flourishes in autumn and summer; conversely, if day-length is shorter than 0.5, we set a range with smaller mean value of K_c and SC_{max} for winter and spring. The equations (Leavesley et al., 1983) can be summarized as:

$$P_{int} = \begin{cases} K_c \times D_c \times A, 0 \le S_0 \le SC_{max} \\ 0\,, \ S_0 < 0 \\ P_{max}, \ S_0 > SC_{max} \end{cases} \quad (13)$$

$$\frac{dS_0}{d_t} = P_{int} \quad (14)$$

$$\begin{cases} K_c = K_{cs};\ SC_{max} = SC_{Smax}, when\ DayL \geq 0.5 \\ K_c = K_{cw};\ SC_{max} = SC_{Wmax}, when\ DayL < 0.5 \end{cases} \quad (15)$$

where D_c is vegetation coverage, which reflects the spatial coverage of vegetation. A is catchment area. Values of D_c and A are from the CAMELS dataset.

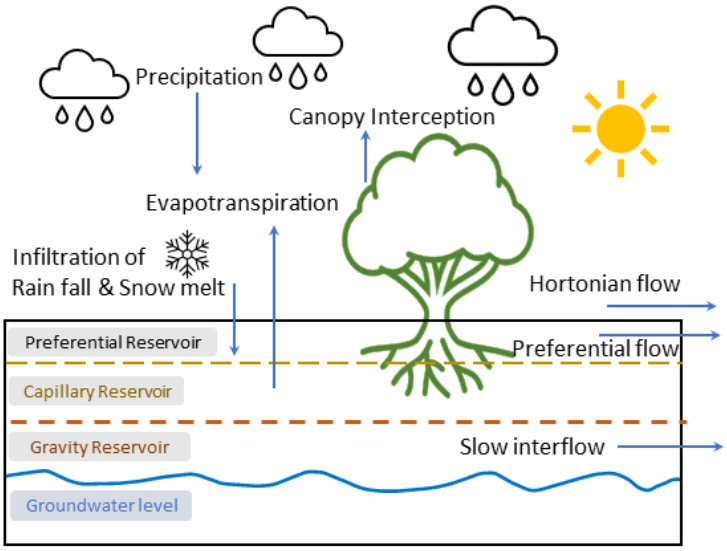

Fig. 4. The main hydrological processes considered in the proposed model.

2.3.2 Snow melt

Snow melt process, a critical factor for hydrological simulation especially in cold regions, has been proven to be a key point for streamflow simulations with DL methods (Broxton et al., 2019). Inspired by the methods of wrapping EXP-HYDRO model (Patil and Stieglitz, 2014) into P-RNN model (Jiang et al., 2020), the snow melt process is implanted into the hybrid model as follows.

First, the precipitation reaches the ground will be divided into snowfall (P_s) and rainfall (P_r), which is controlled by daily temperature (T) and threshold temperature of snowfall (T_{min}):

$$\begin{cases} P_r = P - P_{int}; P_s = 0, when\ T > T_{min} \\ P_s = P - P_{int}; P_r = 0, when\ T \leq T_{min} \end{cases} \quad (16)$$

Then, water from snow melt was calculated as:

$$M = \begin{cases} \min\{S_1, D_f \times (T - T_{max})\}, \ if S_1 > 0 \ and \ T > T_{max} \\ 0, \ otherwise \end{cases} \quad (17)$$

where D_f is thermal degree-day factor, T_{max} is temperature threshold of snowmelt, and S_1 was the snow reservoir, which can be expressed as:

$$\frac{dS_1}{d_t} = P_s - M \quad (18)$$

2.3.3 Soil retention and groundwater flow

Describing the dynamic system of water movement in soil has always been a complex but critical object for groundwater simulation (Banwart et al., 2019; Cai et al., 2021). Indeed, soil structure affects the all-round local hydrological response (Fatichi et al., 2020). What we focused on was the water distribution in soil that meets the catchment-scale water balance requirements. Again, this part is referred to the soil structures introduced in GSFLOW model. Specifically, we only considered the downward vertical flow process between preferential reservoir, capillary reservoir, and gravity reservoir, with the aim of meeting the catchment-scale water balance.

2.3.3.1 Preferential reservoir

When precipitation and snow melt reach the ground, a fraction of infiltration is apportioned to the preferential-flow reservoir to account for fast interflow through large openings in the soil zone near land surface, while the rest of the water, which is generated when the precipitation rate exceeds the infiltration rate of the soil that may not be saturated will be classified as Hortonian flow (Horton, 1933). Eventually, both Hortonian flow and preferential flow contribute to the streamflow of catchment outlet. Referring to the TOPMODEL and P-RNN model, water flow in the preferential-flow bucket can be calculated as:

$$Q_{hor} = (P_r + M) \times (1 - R_{in}) \quad (19)$$

$$Q_{pref} = \begin{cases} 0, & S_2 < 0 \\ (P_r + M) \times R_{in} \times R_{pr} \times e^{-f \cdot (S_{pmax} - S_2)}, & 0 \le S_2 \le S_{pmax} \\ Q_{pmax}, & S_2 > S_{pmax} \end{cases} \quad (20)$$

where R_{in} is the infiltration rate; R_{pr} is the coefficient of fast interflow; f is decay factor; S_2 is preferential reservoir; S_{pmax} is the storage capacity of preferential bucket; and Q_{pmax} is the maximum preferential flow. The values of the parameters are learned during iterative process of self-designed RNN model, which can be expressed as:

$$\frac{dS_2}{d_t} = (P_r + M) \times R_{in} \times R_{pr} \quad (21)$$

2.3.3.2 Capillary reservoir

The capillary reservoir represents water held in the soil by capillary forces between the wilting and field-capacity thresholds. Water is removed from the reservoir by evapotranspiration (Markstrom et al., 2008). Referred to the EXP-HYDRO model and P-RNN model, the PET (Potential Evapotranspiration) is estimated by Hamon's formulation. Therefore, the calculation can be concluded as follows:

$$Q_{cap} = (P_r + M) \times R_{in} \times \left(1 - R_{pr}\right) \quad (22)$$

$$\frac{dS_3}{dt} = Q_{cap} - ET \quad (23)$$

$$ET = \begin{cases} 0, & S_3 < S_{cmin} \\ PET \times \left(\frac{S_3}{S_{cmax}}\right), & 0 < S_3 < S_{cmax} \\ PET, & S_3 > S_{cmax} \end{cases} \quad (24)$$

$$PET = 29.8 \times L_{day} \times \frac{0.611 \cdot e^{17.3 \cdot \frac{T}{T+237.3}}}{T + 237.3} \quad (25)$$

where Q_{cap} is the waterflow into capillary reservoir, S_3 is capillary reservoir, S_{cmin} and S_{cmax} are the minimum and maximum storage capacities of capillary reservoir, L_{day} is day length and T is

local temperature. The main function of capillary reservoir is to calculate the amount of evapotranspiration. The maximum storage capacity represents the field capacity, and the minimum storage capacity represents the minimum storage capacity held by vegetation.

2.3.3.3 Gravity reservoir

The gravity reservoir represents water in the soil zone between field-capacity and saturation thresholds. This reservoir was developed to provide gravity drainage from the soil zone to the unsaturated zone, which will eventually discharge to the groundwater. According to GSFLOW model, slow interflow from the gravity reservoirs represents the perching of water in the soil zone above the water table that can occur because of mineralization near the bottom of the soil zone or when soil develops over fine-grained material. Slow interflow can occur when the water content in the soil zone exceeds the field-capacity threshold. Slow interflow is developed from continuity and an empirical equation (Leavesley et al., 1983). For the lumped hydrological model, we added a decay process for slow interflow referring to TOPMODEL, which can be written as:

$$Q_{slow} = \begin{cases} 0, & S_4 < 0 \\ Q_{gra} \times k_l + Q_{gra}^2 \times k_n \times e^{-f \cdot (S_{gmax} - S_4)}, & 0 \le S_4 \le S_{gmax} \\ Q_{gmax}, & S_4 > S_{gmax} \end{cases} \quad (26)$$

$$Q_{gra} = Q_{cap} - ET \quad (27)$$

$$\frac{dS_4}{d_t} = Q_{slow} \quad (28)$$

where Q_{gra} is the water flows into gravity reservoir; S_4 is gravity reservoir; S_{gmax} is maximum storage capacity of gravity reservoir; k_l is linear coefficient of slow interflow; k_n is non-linear coefficient of slow interflow; and Q_{gmax} is the maximum slow interflow.

Finally, the outputs of the water balance model are expressed as:

$$Q_{stream} = Q_{hor} + Q_{pref} + Q_{slow} \quad (29)$$

$$Q_{gw} = Q_{gra} - Q_{slow} \quad (30)$$

where Q_{stream} is the water flowing into stream at catchment outlet and Q_{gw} is the proportion of water that flows into groundwater at the catchment scale.

Table 1. Pseudocode of the ODE-RNN layer for GWL simulation.

Algorithm: the water balance wrapped GWL-RNN layer

Input forcings: Precipitation $\langle \boldsymbol{P} \rangle$, temperature $\langle \boldsymbol{T} \rangle$ and day length $\langle \boldsymbol{Dayl} \rangle$

Self-learnable parameters:

$\boldsymbol{SC_{max}}$, $\boldsymbol{K_c}$, $\boldsymbol{T_{min}}$, $\boldsymbol{T_{max}}$, $\boldsymbol{D_f}$, $\boldsymbol{R_{in}}$, $\boldsymbol{R_{pr}}$, $\boldsymbol{S_{pmax}}$, $\boldsymbol{Q_{pmax}}$, $\boldsymbol{f}$, $\boldsymbol{S_{capmax}}$, $\boldsymbol{S_{capmin}}$, $\boldsymbol{k_l}$, $\boldsymbol{k_n}$, $\boldsymbol{S_{gmax}}$, $\boldsymbol{Q_{gmax}}$

State initialization: $\boldsymbol{S_0^{(0)} = 0}$, $\boldsymbol{S_1^{(0)} = 0}$, $\boldsymbol{S_2^{(0)} = 0}$, $\boldsymbol{S_3^{(0)} = 0}$ and $\boldsymbol{S_4^{(0)} = 0}$

Calculate PET (mm) via Hamon's formulation Eqs. (24) and (25)

Define step-functions ([$\boldsymbol{P^{(i)}}$, $\boldsymbol{T^{(i)}}$, $\boldsymbol{PET^{(i)}}$], [$\boldsymbol{S_0^{(i-1)}}$, $\boldsymbol{S_1^{(i-1)}}$ $\boldsymbol{S_2^{(i-1)}}$ $\boldsymbol{S_3^{(i-1)}}$ $\boldsymbol{S_4^{(i-1)}}$], self-learnable parameters) via Eq. (13) to Eq. (28)

Calculate $\boldsymbol{P_{int}^i}$, $\boldsymbol{P_s^i}$, $\boldsymbol{P_r^i}$, $\boldsymbol{M^i}$, $\boldsymbol{Q_{cap}^i}$, $\boldsymbol{Q_{pref}^i}$, $\boldsymbol{Q_{hor}^i}$, $\boldsymbol{Q_{gra}^i}$, $\boldsymbol{Q_{slow}^i}$

Calculate $\boldsymbol{S_0^{(i)}}$, $\boldsymbol{S_1^{(i)}}$, $\boldsymbol{S_2^{(i)}}$, $\boldsymbol{S_3^{(i)}}$, $\boldsymbol{S_4^{(i)}}$

Return $\boldsymbol{S_0^{(i)}}$, $\boldsymbol{S_1^{(i)}}$, $\boldsymbol{S_2^{(i)}}$, $\boldsymbol{S_3^{(i)}}$, $\boldsymbol{S_4^{(i)}}$

Run RNN (step-function,[*P*, *T*, *Dayl*, *PET*], [$\boldsymbol{S_0^{(0)}}$, $\boldsymbol{S_1^{(0)}}$, $\boldsymbol{S_2^{(0)}}$, $\boldsymbol{S_3^{(0)}}$, $\boldsymbol{S_4^{(0)}}$]) for sequences of $\boldsymbol{S_0}$, $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, $\boldsymbol{S_3}$, $\boldsymbol{S_4}$

Calculate timeseries $\boldsymbol{Q_{gw}}$ by Eqs. (29) and (30)

Output: the sequence of GWL near the catchment outlet ($\boldsymbol{Q_{gw}}$)

To better illustrate the way of wrapping above equations into WB-Model, we use snow melt process in subsection 2.3.2 as an example. The precipitation that reaches the ground (flux from canopy interception reservoir) and temperature in input (x_t) will be used for initial input for this section, the snow reservoir. Firstly, we set a temperature threshold T_{min} ranging from -3℃ to 0℃ to determine whether precipitation reaching the ground is in form of rainfall or snow (Eq. 16).

Secondly, we set a temperature threshold T_{max} ranging from 0℃ to 3℃ to determine when the snow starts to melt. A thermal degree-day factor D_f ranging from 0 to 5 (mm/day/℃) was set to calculate the snow melt (Eq. 17). The state variable (S_1^t), which represent the snow storage, is updated by Eq. 18 as $S_1^t = S_1^{t-1} + P_s^t - M^t$. In this case, T_{min}, T_{max} and D_f will be determined through the training process. Above process is inserted into the neural network by defining S_1, rainfall, snow and melt with Eq. 16 to Eq. 18. Further, the flux out of snow reservoir (S_1), which is the rainfall and snow melt, will serve as the input for preferential reservoir in subsection 2.3.3.1 to further update S_2 and calculate other fluxes. With similar process, state variables will be updated in order from S_0 to S_4 indicating water flows from precipitation to groundwater. As a result, after updating the five states and nine flux variables, the preliminary groundwater fluctuation would be calculated as Eq. 30. Pseudocode of the ODE-RNN layer for GWL simulation is presented in Table 1. Overall, the process-based groundwater model involves three input daily variables from the CAMELS dataset (Precipitation, averaged Temperature and Day-length), five states variables (S_0, S_1, S_2, S_3 and S_4), nine flux variables (P_{fall}, P_s, P_r, M, Q_{cap}, Q_{pref}, Q_{hor}, Q_{gra} and Q_{slow}), and 16 learnable parameters, which controls the hydrological behaviors (SC_{max}, K_c, T_{min}, T_{max}, D_f, R_{in}, R_{pr}, S_{pmax}, Q_{pmax},f, S_{cmax}, S_{min}, k_l, k_n, S_{gmax}, Q_{gmax}). Again, the proposed WB-Model is a spatially lumped DL model with physical constrains, which adheres strictly to the law of water balance.

2.4 Integrated hybrid framework for GWL simulation

Inspired by the outstanding work of P-RNN integrated hybrid streamflow simulation model (Jiang et al., 2020) and previous work of exploring the usage of DL model for large-scale GWL simulation (Cai et al., 2021), this study proposed a hydrology-aware DL architecture for GWL simulation. As presented in Fig. 5, there were two pipelines for the hybrid model. The first pipeline

was the water balance based RNN model (WB-Model), which wrapped catchments attributes and water balance equations into DLA. The forcing variables for this pipeline were precipitation, day-length, and temperature. The parameters within WB-Model were trained separately to present a preliminary GWL, which was furtherly served as one of the input variables for the second pipeline. The second pipeline was a sequence-to-sequence DL model, which maps the meteorological sequences to GWL sequence. The input variables of the second pipeline included the preliminary GWL simulated by WB-Model and the other six dynamic forcings provided by the CAMELS dataset. The objective of WB-Model was to identify geo-hydrological information from observations of the external world and reorganize them in form of providing simulated preliminary GWL. In other words, WB-Model was used to facilitate the prevalent DL model. The core function of WB-Model was to combine water balance equations and DL algorithm to capture the fluctuation characteristics of GWL that response to driven factors. As a result, the preliminary GWL simulated by WB-Model was served as an important input forcing to strengthen the training process of hybrid model. Considering the ability of handling lagged effect from hydrological signals (Feng et al., 2020) and a faster computing speed than RNN-based model (Jiang et al., 2020), a two-layer 1D-CNN model was chosen as the main DLA for simulating the GWL. After a large number of preliminary experiments considering the trade-off between increasing modelling accuracy and reducing the complexity of the model structure, the first layer applied 8 kernel filters, each with a length of 15. The length of 15 symbolized that the influence on the current hydrological response could be traced back to 15 days ago. From a hydrological viewpoint, tracing the influences on the current hydrological response back to 10 days ago is a common strategy in data-driven hydrological models and has been proven to be successful for streamflow simulation with similar method (Jiang et al., 2020). Since our simulation target is GWL for phreatic aquifer, the

hydrological response time will be a little longer than that of streamflow simulation. The set of 15-day hydrological response time was based on a large number of preliminary experiments. Nevertheless, we cannot ensure that 15-day hydrological response time is the most accurate setting for each specific catchment. However, this setting is overall feasible for GWL simulations of these 91 catchments and is in line with common strategy in data-driven hydrological models. The second layer used 1 convolution kernel filter for analyzing the output of the first layer and providing the final results. Zero-padding strategy was adopted for ensuring the output had the same sequence length as the input sequence after filtering by kernel filters.

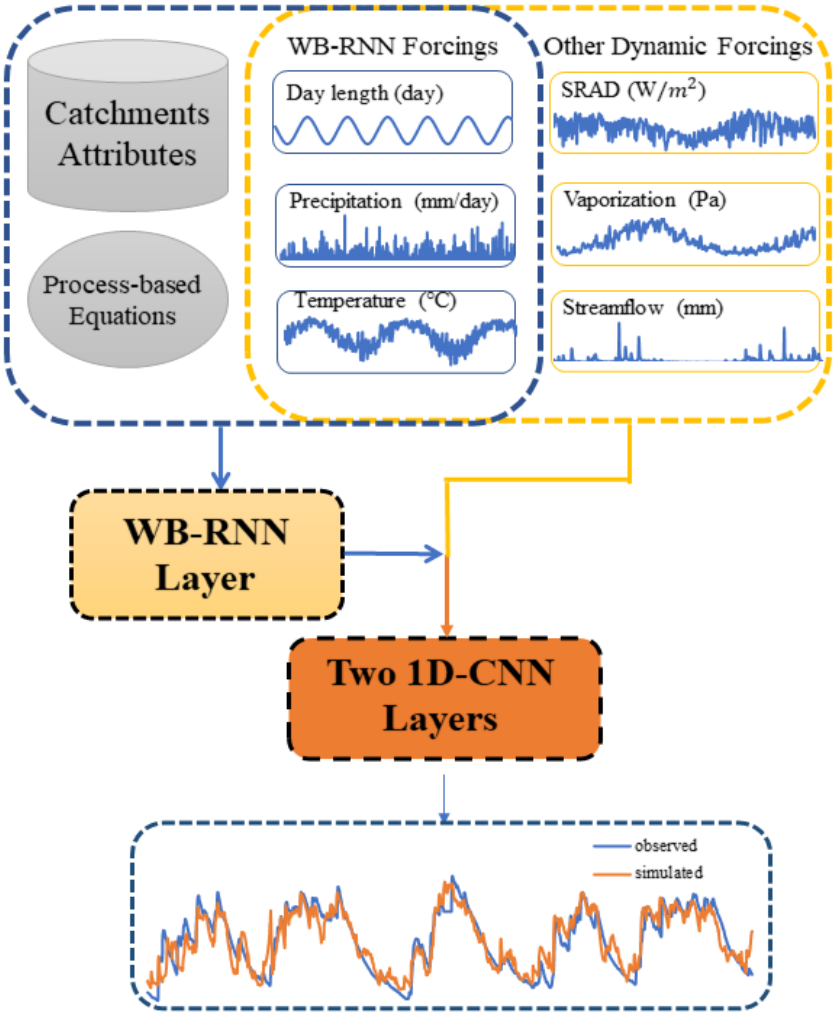

Fig. 5. The proposed generic architecture that explicitly embedding water balance constrained dynamic behaviors into DL models.

The performance of the model was evaluated with Nash-Sutcliffe coefficient of Efficiency (NSE) value (Nash and Sutcliffe, 1970) as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_s - Q_o)^2}{\sum_{i=1}^{n}(Q_O - \overline{Q_O})^2} \tag{31}$$

where Q_s, Q_o and $\overline{Q_O}$ are the simulated, observed and mean observed GWL. NSE is widely used for model assessment, which ranges from $-\infty$ to 1. The closer NSE is to 1, the better the simulation result is. The NSE value was also set as loss function for all the DL models in this study. 80% of each timeseries was divided as training set while the rest 20% of the data was used as testing set. The epochs number was set as 400 and the learning rate was 0.01. The number of parameters in hybrid model is 593 including 16 self-designed parameters in WB-Model, 568 parameters in first 1-D CNN layer and 9 parameters in the second 1-D CNN layer. All the models and equations introduced in this study were coded with Python 3.6 with Keras as coding API and Tensorflow 1.4 as the DL backend with no GPU requirement.

Generation ability (GA) is an important evaluation criterion for data-driven model (Chen et al., 2020). It reflects the ability of DL model, which is well trained, to digest new data and make accurate predictions. The calculation method is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_s - O_O)^2} \tag{32}$$

$$GA = \frac{RMSE_{pred}}{RMSE_{train}} \tag{33}$$

where O_s and O_o are the simulated and observed GWL; *RMSE* refers to Root Mean Square Error, which measures the average magnitude of the error between model simulations and observations; and *GA* is calculated as the ratio of the *RMSE* values between predicting process and training

process (Yoon et al., 2011). If the data-driven model simulates GWL perfectly, the *GA* values will be unity. Otherwise, if the model is over trained or overfitting, the *GA* values will exceed unity. In this case, the model would not give accurate predictions even though it is able to provide accurate fittings for the training data. The *GA* values will be less than unity if the model is under trained, which is also called as underfitting.

3 Results and discussion

3.1 Comparisons of GWL simulation results

3.1.1 Comparisons of model performances in specific catchments

As representatives of the 91 catchments, three catchments (HUC: 04216418, HUC: 03015500 and HUC: 03026500) were chosen for presenting, describing, and analyzing the simulations in detail. Fig. 6 presented the GWL forecasting performances with four independent DL strategies, from top to bottom of the subplots are Hybrid model, 1D-CNN model, WB-Model (the self-designed RNN model), and GRU model. WB-Model integrated water balance constrains into RNN, which served as a major optimization for hybrid DL model while the 1D-CNN model and GRU model are pure DL model. Consequently, what concerned us the most was how much the simulation performance of hybrid model was improved compared with the 1D-CNN model and GRU model. The reason why we chose these three catchments as representatives was that the simulation performances of the 91 catchments could be roughly divided into three categories, which corresponded to the three catchments shown in Fig. 6.

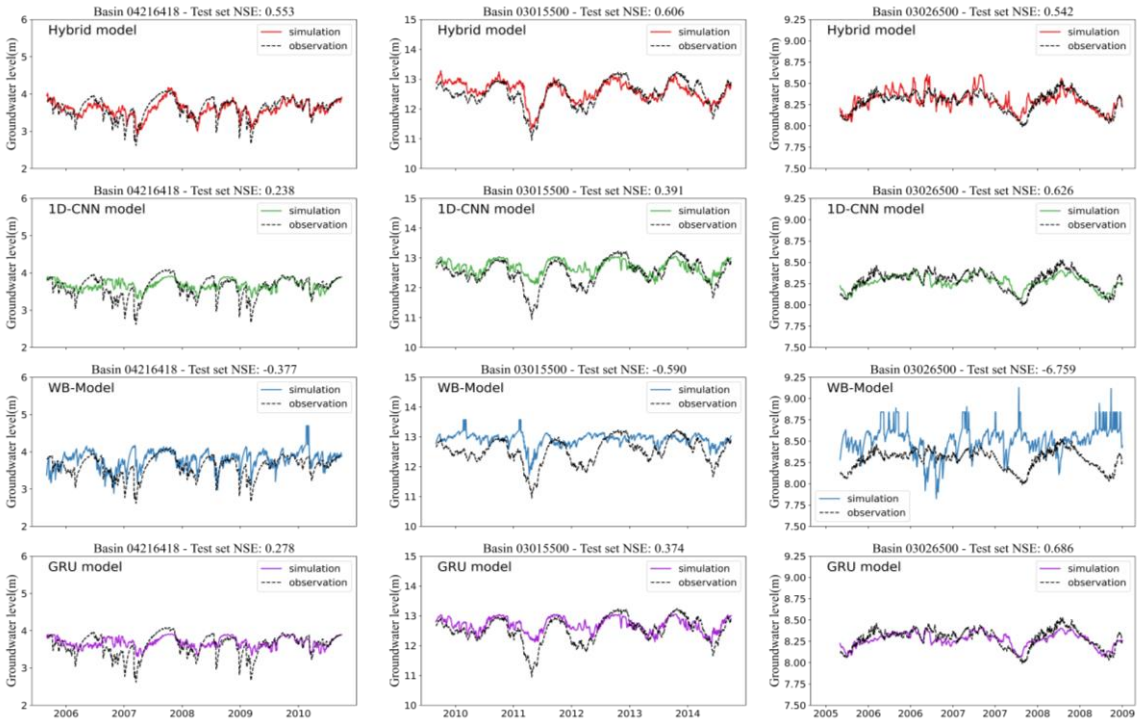

Fig. 6. Representatives of simulation results by four DL strategies introduced in this study.

The simulation results of HUC: 04216418 represented the catchments that with an overall simulation improvement when combining water balance constrains into common DL models. As shown in the first column of Fig. 6, the pure DL strategies, 1D-CNN (NSE=0.238) and GRU (NSE=0.278) model, had similar simulations effect while the hybrid model (NSE=0.553) outperformed significantly than the pure DL models. Such improvement was majorly benefit from the accuracy of simulating the GWL fluctuance pattern by the WB-Model (NSE=-0.377). Admittedly, based on the evaluation criteria of hydrological model, the performance of WB-Model itself were not satisfied enough for an alternative of traditional process-based model. However, although there were certain differences in specific values, the variation trend and fluctuation patterns of the simulation results were consistent with the observations. As a result, the preliminary result from WB-Model, served as an important input forcing, strengthened the training process of

hybrid model, which lead to a higher accuracy than pure DL models. More specifically, in plain language, common DLAs possessed the ability to filter out critical information that mostly related to the simulation target by adjusting the intrinsic parameters, such as weights and bias. In this case, comparing to the other two representatives, preliminary GWL simulated by WB-Model provided the most useful sequence for mapping the observations because the fluctuance pattern simulated by WB-Model was most similar to that of the observed GWL.

The second column of Fig. 6, HUC: 03015500, presented the representative of catchments with another type of simulation performance. In general, despite the NSE improvements may not be as obvious as the catchments in the first category, hybrid models (NSE=0.606) outperformed the pure DL models (NSE_CNN=0.391 and NSE_GRU=0.374) in these catchments for better simulating the GWL fluctuance under extreme conditions. Comparing to the first category, the overall simulation performances of WB-Model in this category were slightly inferior, which led to a less obvious promotions of NSE values. As the simulation of catchment HUC: 03015500 illustrated, there was an obvious increasement of GWL during 2011 to 2012, which was resulted from a large amount of snow melting and rainfall as precipitation. As shown in the 1D-CNN model and GRU model, the pure DL models failed to capture such precipitation pattern because precipitation characteristics, especially the extreme precipitation conditions, had certain probability of being diluted in a variety of input variables. During the training process, DL model would consider the extreme precipitation events as outliers especially when the other inputs were sufficient to establish a good liner or nonlinear mapping relationship with the object. This could be inferred from the fact that pure DL models had already presented good overall simulation performances. Under these circumstances, WB-Model captured the pattern of extreme precipitation events when providing the preliminary GWL, which emphasized the importance of

the extreme points and reinforced the hybrid model to learn such patterns. Consequently, the simulation results of hybrid model were intuitively improved because of better matchings of extreme GWL.

Table 2. Summary of the simulation results of testing set by hybrid, 1D-CNN, and GRU model across 91 catchments in the 10 major watersheds.

Watersheds (Region)	Basin count	1D-CNN		GRU		Hybrid Model		Improvement	
		Median	Mean	Median	Mean	Median	Mean	Mean	Count
01 New England	5	0.275	0.282	0.277	0.291	0.278	0.330	-0.008	3
02 Mid Atlantic	20	0.299	0.305	0.313	0.326	0.393	0.405	0.125	17
03 South Atlantic-Gulf	15	0.256	0.326	0.260	0.313	0.356	0.361	0.254	13
04 Great Lakes	5	0.365	0.348	0.373	0.347	0.528	0.578	0.231	5
05 Ohio	16	0.329	0.349	0.338	0.358	0.388	0.459	0.036	10
06 Tennessee	6	0.243	0.210	0.300	0.293	0.334	0.334	0.106	6
07 Upper Mississippi	8	0.298	0.354	0.443	0.275	0.475	0.475	0.097	7
08 Lower Mississippi	3	0.603	0.533	0.663	0.559	0.531	0.531	-0.013	2
10 Missouri	7	0.192	0.288	0.434	0.344	0.224	0.224	0.127	4
11 Arkansas-White-Red	6	0.314	0.401	0.432	0.548	0.409	0.409	0.244	4
ALL	91	0.279	0.253	0.313	0.278	0.408	0.401	0.129	71

The last column of Fig. 6, HUC: 03026500, was presented as a representative of catchments with negative effects when integrating WB-Model. Intuitively, preliminary GWL simulated by WB-Model failed to match the fluctuance of observed GWL (NSE=-6.759). Conversely, the pure DL models showed satisfying simulation results (NSE_CNN=0.626 and NSE_GRU=0.686). The reduction of hybrid model performance (NSE=0.542) resulted from the failure of WB-Model. Specifically, for catchment HUC: 03026500, the GWL simulated by hybrid model showed a larger fluctuation range than observed GWL, which was consistent with simulation from WB-Model. There were three potential reasons: 1) the groundwater observation well and the corresponding

streamflow observation station located in different drainage areas. As mentioned in subsection 2.1, although the groundwater data we collected were from the observation well that was closest to the streamflow observation station at the catchment outlet, we cannot guarantee that the GWL was driven by the precipitation from the streamflow catchment because the catchments of groundwater and surface flow might not be geographically coincident. In this case, the WB-Model driven by the precipitation from the CAMELS dataset would inevitably lead to failed GWL simulation results because the fluctuation of GWL was driven by a different precipitation input. 2) The water balance algorithm was not applicable for local GWL simulation. The essences of the WB-Model and hybrid model were physical constrained DL models for lumped GWL simulation, which meant the detailed geological information related to groundwater flux were not collected or considered in these models. This would lead to simulation mismatches if discharge and recharge of local groundwater were not majorly from the precipitation but from other sources such as confined aquifer or lakes nearby. In this case, the WB-Model would not provide a rather precise simulation result even if the precipitation was one of the driven forces for the local groundwater fluctuance. 3) The ranges of learnable parameters were not set properly. This was also resulted from the lack of geophysical information to determine the range of self-designed parameters especially the soil related parameters. Overall, the influence from the third problem could be minimized by multiple tests while the first and second problems might be the most likely causes of the failures of WB-Model and hybrid model for GWL simulations in several catchments.

3.1.2 Comparisons of overall accuracy between hybrid model and pure DL models

As shown in Table 2, 1D-CNN, GRU and water balance constrained hybrid model were applied to simulate GWL of 91 catchments located in 10 major watersheds of the middle eastern CONUS. More specifically, the median NSE values of the simulation results by using GRU model

was 0.313 while the average NSE was 0.278, which was consistent with the NSE range of former simulations in similar areas by applying GRU algorithm (Cai et al., 2021). As an advanced DLA focused on timeseries simulation, GRU model served as a group of controlled experiments in this study for physical constrained hybrid model. In the meanwhile, the median NSE value of simulation by using 1D-CNN model was 0.297 and the average NSE was 0.253. In views of the NSE values, the simulation performance of the two pure DL models were not much different, but due to the complexity of model structure and the difference in emphasis of simulation (GRU model focuses on timeseries simulation and CNN model focuses on image processing), the overall result of 1D-CNN model simulation would be slightly inferior to that of GRU model for GWL simulations. Benefited from the WB-Model, the overall simulation performance was significantly better than that of 1D-CNN and GRU model: the median NSE was 0.408 and the average value was 0.401. The NSE distribution was shown in Fig. 7(b).

Since the final algorithm of hybrid model was a two-layers 1D-CNN model, comparing to the pure 1D-CNN model, the preliminary GWL simulated by the WB-Model was the key factor for improving the performances of hybrid model. With the aim of testing the adaptability of hybrid model, we counted the number of catchments where the hybrid model outperformed the 1D-CNN model for GWL simulation. As shown in the last two columns of Table 2, comparing with the 1D-CNN model, 71 out of 91 catchments presented simulation improvement when applying the hybrid model. Furthermore, for the 10 major watersheds, comparing to the 1D-CNN model, average NSE improvements while using hybrid model were positive except for the New England and Lower Mississippi regions. However, the number of catchments used for experiments was so small that the average NSE improvement would be seriously affected by a single catchment with a major failure of hybrid model. Moreover, more than half of the catchments in New England and Lower

Mississippi region presented positive improvement when using hybrid model than using 1D-CNN model. Optimizing effect of hybrid model was more intuitively explained in Figs. 7(a) and 7(c). Distribution of catchments where hybrid model outperformed the 1D-CNN model were shown in Fig. 7(d).

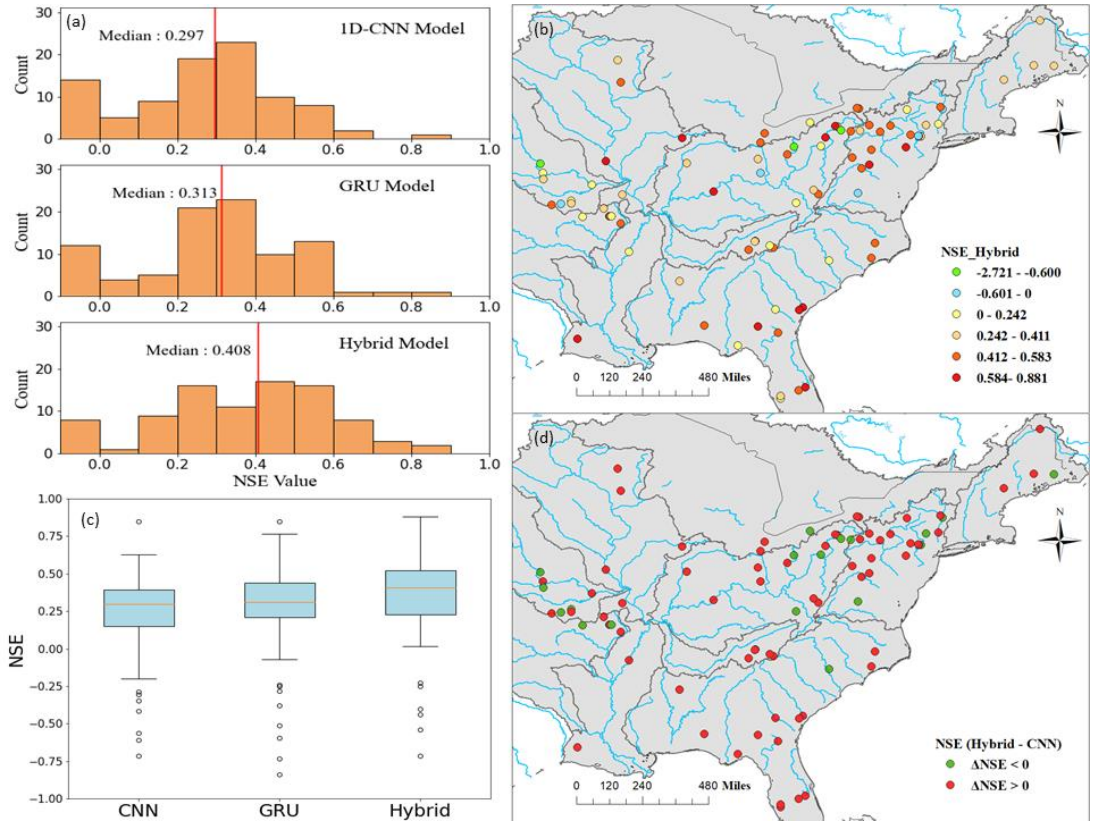

Fig. 7. (a) Histogram for comparison of simulation result from hybrid, 1D-CNN and GRU model; (b) Distribution of NSE values for simulation results of hybrid model; (c) Box plot for comparison of simulation result from hybrid, 1D-CNN and GRU model; (d) Distribution of catchments where hybrid model outperformed 1D-CNN model.

Lumped hybrid DL model for streamflow simulation based on snow melting processes had been tested to be successful for catchments in high latitudes (Jiang et al., 2020). In this study, in addition to the snow melting process, groundwater-related soil water flow processes were also properly embedded into the WB-Model. As a result, the hybrid model outperformed the pure DL

model not only in the snow-domain areas but also basins in middle and low latitudes. Preliminary GWL simulated by WB-Model improved the noise tolerance of hybrid model to groundwater-irrelevant input features and strengthens the learning ability of the hybrid model to GWL changes. Therefore, comparing to the pure DL models, hybrid model possessed better adaptability and understanding of different hydrometeorological and geophysical conditions with the constrains of water balance equations. In conclusion, comparing to prevalent pure DL models, hybrid model exhibited significantly higher overall accuracy for GWL simulations.

3.2 Comparisons of model generalization ability

The generalization ability was evaluated by Eq. 33. If the model focused on the training process rather than a general system, the GA value would be higher. This meant the model will have overfitting problems. The higher the GA values were, the weaker the generalization ability the model processed. Fig. 8 presented the distributions of the GA values for the four DL models in form of boxplot. The average GA values of hybrid, WB-Model, CNN and GRU model were 1.269, 1.124, 1.431 and 1.396, respectively. This indicated the developed hybrid model presented a better generalization ability than that of CNN and GRU model and WB-Model presented the best overall generalization ability among the four DL models. It should be noted that WB-Model was a group of ODE equations coded with the language used in recurrent neural networks. This meant that WB-Model was essentially a physic guiding model. For WB-Model, the water balance equations, served as a backbone of training and predicting stages, were universal theorems. This was the main reason why WB-Model presented the best generalization ability. Furthermore, the result was consistent with conclusions of the previous study, which proved that physic based numerical models tended to present better generalization abilities than that of pure machine

learning models (Chen et al., 2020). In conclusion, benefited from WB-Model, hybrid model presented superiority than pure DL models in terms of generalization ability.

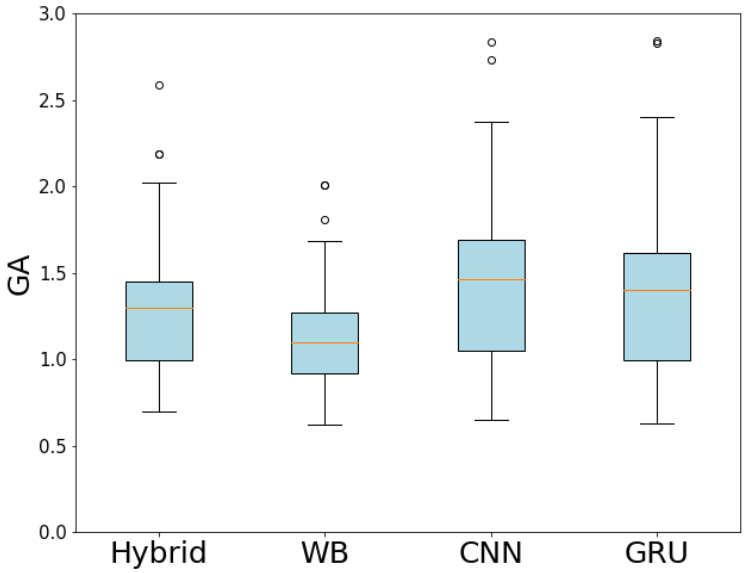

Fig. 8. Comparison of the GA values between Hybrid, WB-Model, CNN and GRU model.

3.3 Robustness of the proposed model

In general, in addition to the advanced core algorithms, an important factor affecting the effect of DL model was the amount of sample data as well as the potential relationship between input data and target data. An excellent DL models should be equipped with the ability to ensure its own stability while satisfying the accurate simulation performance while reducing the effect from the interference information as much as possible (Su et al., 2018). To demonstrate the reliability and stability of the proposed model, we chose the catchment HUC: 03182500 for scenario tests. We evaluated the GWL simulation results of hybrid model with six combinations of input features. More input features signified a larger amount of input data, which could be either useful information or noisy for simulations of targets. In this study, since the WB-Model needed three inputs (precipitation, temperature, and day length) for the self-designed RNN model, these three

variables were the minimum input requirement. As shown in Fig. 9, with the increasing number of input features, NSE values of the hybrid model also presented an increasing trend, from 0.564 to 0.650. It was worth mentioning that although the NSE values of the three inputs strategy was the smallest, the value of NSE=0.564 was a fine result for GWL simulations. Meanwhile, in perspective of the NSE values, the difference of simulation performances between the six scenarios was relatively small, which proved that the hybrid model processing a good stability. Furthermore, it could be inferred that, comparing to SRAD and vapor, applying streamflow data as one of the input features could better improve the accuracy of the hybrid model, which implied that the feature of an important hydrological phenomenon, flow exchange between streamflow and groundwater through baseflow, was captured from the perspective of hybrid DL model. This reflected that DL models, which had always been criticized as black-box models without any physical meanings, could also express some natural processed implicitly (Jiang et al., 2022).

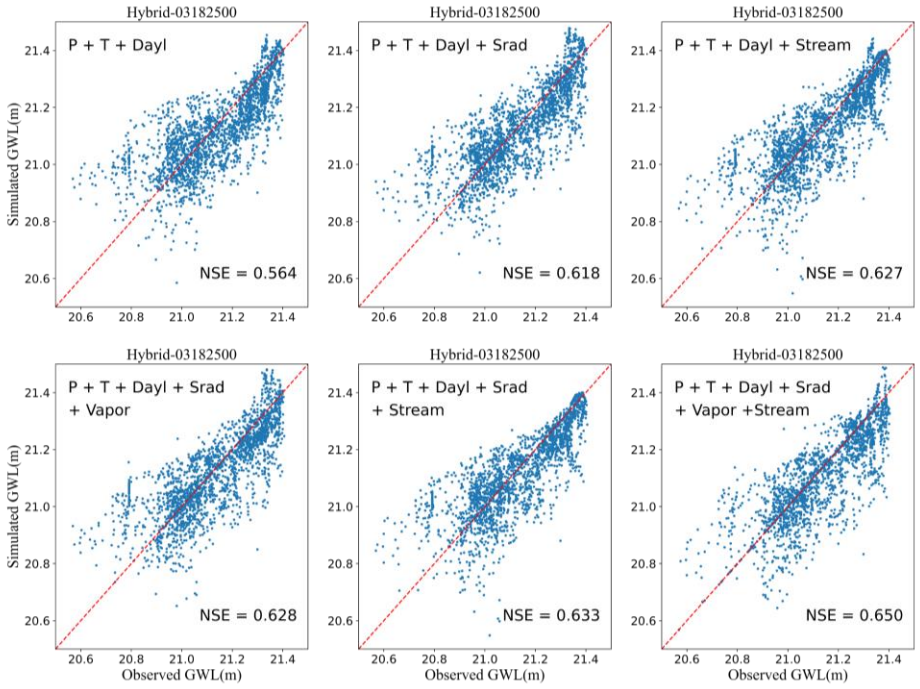

Fig. 9. Comparisons of hybrid model performances with different input strategies.

Fig. 10 presented the comparisons of the stability of physical constrained hybrid model and 1D-CNN model as representative of pure DL model through the NSE values. We adopted the data from catchment HUC: 03182500 for testing experiments to compare the performance of the model under same input feature strategies. The result showed that hybrid model outperformed the pure DL model for every input strategy and the NSE value of hybrid model presented a smaller fluctuation range (from 0.564 to 0.651) than that of pure DL model (from 0.417 to 0.517), which implied the hybrid model performed more stably than pure DL model.

The reason for this phenomenon was that the WB-Model in the hybrid model could adjust the values of the internal learnable parameters according to the regional characteristics, so that the preliminarily simulated GWL could reach a certain accuracy. Then, as an input, the preliminarily simulated GWL could help the DL model better understand the change characteristics of the GWL to be simulated, so that the performance the hybrid model was obviously better than that of the pure DL model. In conclusion, the physical constrained mechanism made the stability of the hybrid model better than the pure DL model, which was another outstanding advantage in addition to the higher accuracy of the simulation results of the hybrid model.

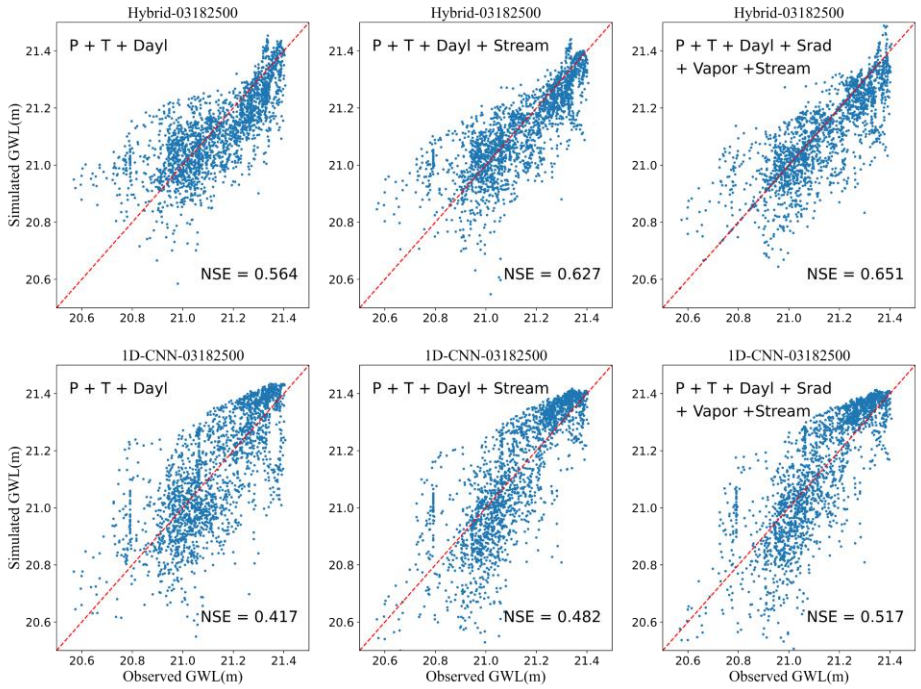

Fig. 10. Comparison of simulation performances between hybrid and 1D-CNN model with different input strategies.

3.4 A rethink of relationships between simple RNN, WB-Model and GRU.

It is worth clarifying that simple RNN, WB-Model and GRU are based on a general concept of RNN, which is a family of neural networks with recurrent cells. In general, simple RNN model does not possesses the long-term memory ability because of the potential gradient vanishing and explosion problem when updating the hidden states during the training process. The main difference between simple RNNs and GRUs (or LSTMs) is the latter introduces a so-called gating mechanism into the recurrent cell to control the information flow, to address the gradient vanishing and explosion problem. Likewise, in our WB-Model, we also modified the recurrent cells of the general RNN models, where the nodes were connected using physical equations rather than the usual perceptions. More specifically, GRU introduces reset gate and update gate to determine

whether the previous information involved in the cell states is added or discarded. In comparison, in the WB-Model, the information flow is controlled using the thresholds as in hydrological models. For example, if the temperature is larger than a threshold, the rainfall information will become a part of the soil state; otherwise, it will be added to the snow state. Further, although the specific values of these parameters are determined during the training process of DL model, the essence of the WB-Model is a group of ODE equations (presented by TOMODEL, GSFLOW, and other hydrological models) coded with the language used in neural networks, which makes the WB-Model naturally have long-term memory in the form of hydrological states (e.g., snowpack and preferential flow pack). In conclusion, the GRU and WB-Model model have taken distinct ways to reform the recurrent cells of general RNN models, which could be considered as two optimizing strategies of simple RNN model.

3.5 Limitations and potential improvements

Comparing with pure DL models, the proposed hybrid model for cross-region GWL simulations at catchment scale is significantly improved in terms of the accuracy, generality, robustness and physical meaning of models, but there are some obvious limitations and potential improvements for the hybrid model.

The first limitation is that the process of adjusting self-designed parameters could be cumbersome. As introduced in former subsections, there are 16 learnable parameters designed in the WB-Model that control the hydrological behaviors. The ranges of these parameters need to be adjusted manually until the hybrid model produces the most accurate results. For example, the ranges of hydrometeorological parameters that introduced from former models (Jiang et al., 2020; Patil and Stieglitz, 2014), such as T_{min}, T_{max}, f, etc., were set for all catchments while other parameters, especially the geological related and storage related parameters introduced in this

study, such as S_{pmax}, Q_{pmax}, S_{cmax}, S_{min}, etc., should be adjust for specific catchments to reach to best simulation performance. Consequently, the parameters calibration process of hybrid model for some catchments could cost lots of time to achieve the best simulation results.

The second limitation is that due to the lack of detailed geographic information, the lumped GWL model cannot be applicable to all basins. Supporting by the universality of water balance constrains and the powerful computing ability of DLA, our expectation of this hybrid model is that the model can simulate GWL at any point in the basin accurately if the hydrometeorological data is sufficient and the historical GWL observation is provided. This is because we believe the hybrid model can perceive the geological information to give accurate GWL simulations by adjusting learnable parameters set of WB-Model. However, in this study, the hybrid model failed to improve the simulation performances for 20 out of 91 catchments. This is because the lack of detailed geographic information makes some groundwater recharge and discharge processes, such as groundwater recharge from surrounding confined aquifers, impossible to be reflected through single water balance constrains. Therefore, the proposed hybrid model cannot be applied to the simulation of GWL in all catchments, especially in those with complex groundwater recharge and discharge process. This is also reflected in the simulation performance of WB-Model: although WB-Model can simulate the fluctuation state of GWL, most of the simulation results are not accurate, and accurate simulation results are obtained by embedding the preliminary result into another DL structure.

The third limitation is the black box nature of proposed hybrid model remains unrevealed. The main merit of adding water balance ODEs into the DL architecture lies in improving the awareness of hydrological knowledge of hybrid model by guiding WB-Model to capture the fluctuation patterns of GWL. However, adding physics is not enough to reduce the black box

nature of DL model (Barredo Arrieta et al., 2020). Despite we have successfully combined DL with physical constrains, making the hybrid model present extremely powerful forecasting ability, we have not found further interpretable physical meaning from the proposed hybrid model. Especially, for the two-layer 1D-CNN model in the hybrid model, which only provides the outstand non-linear fitting ability, the fitting process remains uninterpreted from hydrological perspective in this study.

The fourth limitation is that the two- or three-dimensional processes of groundwater flow are not considered in this study. In general, the study of groundwater simulations is always related with two or three-dimensional groundwater flow with specific boundary conditions and governing equations. Moreover, detailed groundwater simulation is usually focusing on find numerical or analytical solutions of two or three-dimensional PDEs related to groundwater by numerical method or analytical method (Liang et al., 2018). The hybrid model we proposed is based on the combination of DLA and vertical one-dimensional water balance equation in catchment scale. Although it is a breakthrough to the traditional research methods, it cannot describe the two and three-dimensional groundwater flow processes, which might make it not delicate or convincing enough from the perspective of traditional groundwater flow simulation community. In other words, we embedded physical constraints to improve the simulation performance of DL model, but the physical constraints are not detailed enough compared with traditional methods. More specifically, although water balance constraints with 16 self-learnable parameters are well implanted into WB-model, this number is far less than prevalent DL model, such as 1D-CNN model, which in our case, has 568 parameters in first 1D-CNN layer and 9 parameters in the second 1D-CNN layer. Parsimonious DL model does not ensure a learned model will capture all important information in the data sensed about the external world (Ma et al., 2022). As a result, lacking

complexity is a main reason why WB-Model cannot outperform the GRU model in terms of the accuracy of GWL simulation.

Based on the limitations mentioned above, we believe that this study has potential for further improvements. Firstly, more detailed groundwater-related processes, such as human activities, could be implanted into WB-Model. The proposed hybrid model could be more flexible if multiple potential processes were wrapped into WB-Model for options. In that case, the hybrid model would decide which processes are suitable to be considered for specific catchments by adjusting the parameters and super parameters of WB-Model. This would improve the adaptability of hybrid model greatly if the groundwater-related processes could be considered comprehensively. Secondly, cutting-edge DL interpretive methods, such as expected gradients (EG) (Erion et al., 2021), could be applied to decipher the machine-captured patterns and inner workings of the hybrid model. The main purpose of such study will be revealing the black-box process of the proposed hybrid model, so that the hydrological cognition from the perspective of machine learning models under physical constraints could be obtained to facilitate our improving our understanding for specific hydrological processes. Thirdly, the threshold mechanism of hydrological models in our WB-Model framework might be further improved by the idea of gating mechanism from GRU or LSTM model. Lastly, to optimize the physical constrains from one-dimensional ODEs to higher dimensional PDEs. The difficulty of groundwater simulation is much higher than that of surface water simulation because of its complex two and three-dimensional flow characteristics (Wang et al., 2021). Limited by the lack of geophysical data, the lumped hybrid model based on water balance mechanism has not maximized the power of integrating DLAs and physical constrains. Therefore, the combination of PDEs-governed distributed groundwater flow processes and DL model will be an important research direction in the future. Regardless, the proposed hybrid model

772 is already the most novel model in terms of predicting GWL with time series hydrometeorological
773 inputs.

774 **4 Conclusions**

775 In this study, a lumped hybrid groundwater model with water balance equations as physical
776 constraint and DL methods as core algorithm is proposed to simulate the fluctuation of GWL. In
777 the hybrid model, water balance guided OEDs were wrapped into a WB-Model by self-designing
778 the specific algorithms of RNN model. We tested the model with the CAMELS dataset from 91
779 catchments located in the middle eastern CONUS, and two pure DL models, 1D-CNN and GRU,
780 were established for comparison of simulation performance. The main findings of this study are
781 summarized as follows:

782 First, the hybrid model presented high accuracy of simulating the fluctuation of GWL without
783 using detailed hydrogeological information of the catchments. The preliminary GWL simulated
784 from the WB-Model enhances the learning ability of hybrid model. Consequently, the physics
785 constrained DL model outperformed the pure DL models significantly in 71 out of 91 catchments
786 in this study. Moreover, comparing with traditional distributed GWL simulation models, DLA
787 reduces the cost of data as well as the difficulty of model setup while still provides accurate
788 simulation results in perspective of traditional standards, which makes the hybrid model more
789 suitable for GWL predictions in less gauged or ungauged basins.

790 Second, the self-designed RNN model with water balance constraints proposed in this study
791 embeds the main groundwater-related water balance formulas, which are referred to the traditional
792 distributed hydrological models, into the recurrent neutral networks. The specific values of the
793 parameters in water balance related formula and water storages in designed reservoirs are
794 determined by the hybrid model through iterative algorithms, activation functions and loss

functions of DLAs. Consequently, this equips our hybrid model with the ability to learn the groundwater-related water allocation processes at the catchment scale. In conclusion, the physical constrained hybrid model presents better adaptability and generalization ability comparing with the pure DL models.

Last, compared with the pure DL model, the hybrid DL model proposed has better robustness. This is reflected from the fact that the hybrid models outperformed pure DL model with different strategies of input features. More specifically, the NSE values of simulations from hybrid model have higher values and lower fluctuation range. The reason is that the preliminary simulation results provided by the physically constrained WB-Model strengthen the learning ability of DLA for groundwater fluctuation characteristics, so that the hybrid model can be affected by potential noise data as little as possible.

This study shows the superiority and powerful simulation ability of DL model based on physical constraints. We have abandoned the general idea of building traditional groundwater models: solving two-dimensional or three-dimensional groundwater flow problems through the iterative method of distributed models or finding analytical solutions. Instead, we embedded water balance equations into DLA for regional GWL simulation. As a result, this hybrid model presented great accuracy, adaptability, generalization ability, and robustness even without detailed geological data of catchments, which demonstrated the possibility of application of proposed model for GWL simulation in ungauged or lack of gauged catchments. Although there are many limitations and potential improvements for the proposed model, we believe that the general performance of the proposed model would increase trust in data-driven approaches on hydrological modellings especially when physical constraints related to hydrological sciences are integrated with DLAs.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (51909117) and Natural Science Foundation of Shenzhen (JCYJ20210324105014039).

Author contributions

Conceptualization: Haiyun Shi

Methodology: Hejiang Cai, Haiyun Shi

Formal analysis: Hejiang Cai

Validation: Hejiang Cai, Zhaoqiang Zhou, Shijie Jiang

Software: Hejiang Cai, Zhaoqiang Zhou

Supervision: Haiyun Shi, Suning Liu, Vladan Babovic

Writing - original draft: Hejiang Cai

Writing - review & editing: Haiyun Shi, Suning Liu, Vladan Babovic, Shijie Jiang

Funding Acquisition: Haiyun Shi

Data availability and coding statement

The groundwater level data are openly available at https://waterdata.usgs.gov/nwis/gw, and the CAMELS dataset is freely available at https://ncar.ucar.edu.

We gratefully acknowledge the use of the dataset provided by Addor et al. (2017) and Newman et al. (2015) and the codes of P-RNN model publicly provided by Jiang et al. (2020).

References

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark (2017), The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*, *21*(10), 5293-5313.

Ayzel, G., Heistermann, M., 2021. The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: a case study for six basins from the CAMELS dataset. Comput. Geosci. 149 https://doi.org/10.1016/j.cageo.2021.104708.

Babovic, V. (2009), Introducing knowledge into learning based on genetic programming, *Journal of Hydroinformatics*, *11*(3-4), 181-193.

Babovic, V., and M. Keijzer (2002), Rainfall runoff modelling based on genetic programming, *Nordic Hydrology*, *33*(5), 331-346.

Banwart, S. A., N. P. Nikolaidis, Y.-G. Zhu, C. L. Peacock, and D. L. Sparks (2019), Soil Functions: Connecting Earth's Critical Zone, *Annual Review of Earth and Planetary Sciences*, *47*(1), 333-359.

Barredo Arrieta, A. et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58: 82-115.

Beven, K. J., and H. L. Cloke (2012), Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al, *Water Resources Research*, *48*(1).

Broxton, P. D., W. J. D. Leeuwen, and J. A. Biederman (2019), Improving Snow Water Equivalent Maps With Machine Learning of Snow Survey and Lidar Measurements, *Water Resources Research*, *55*(5), 3739-3757.

Butcher, J. C. (2000). Numerical methods for ordinary differential equations in the 20th century. Journal of Computational and Applied Mathematics, 125(1-2), 1-29.

Cai, H., H. Shi, S. Liu, and V. Babovic (2021), Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States, *Journal of Hydrology: Regional Studies*, *37*.

Chadalawada, J., V. Havlicek, and V. Babovic (2017), A Genetic Programming Approach to System Identification of Rainfall-Runoff Models, *Water Resources Management*, *31*(12), 3975-3992.

Chadalawada, J., H. M. V. V. Herath, and V. Babovic (2020), Hydrologically Informed Machine Learning for Rainfall□Runoff Modeling: A Genetic Programming□Based Toolkit for Automatic Model Induction, *Water Resources Research*, *56*(4).

Chen, C., W. He, H. Zhou, Y. Xue, and M. Zhu (2020), A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China, Sci Rep, 10(1), 3904.

Clark, M. P., et al. (2015a), Improving the representation of hydrologic processes in Earth System Models, *Water Resources Research*, *51*(8), 5929-5956.

Clark, M. P., et al. (2015b), A unified approach for process□based hydrologic modeling: 2. Model implementation and case studies, *Water Resources Research*, *51*(4), 2515-2542.

Cuthbert, M. O., T. Gleeson, N. Moosdorf, K. M. Befus, A. Schneider, J. Hartmann, and B. Lehner (2019), Global patterns and dynamics of climate–groundwater interactions, *Nature Climate Change*, *9*(2), 137-141.

Dunkerley, D. L., and T. L. Booth (1999), Plant canopy interception of rainfall and its significance in a banded landscape, arid western New South Wales, Australia, *Water Resources Research*, *35*(5), 1581-1586.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., & Lee, S. I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature Machine Intelligence, 3(7), 620–631.

Fan, Y., H. Li, and G. Miguez-Macho (2013), Global patterns of groundwater table depth, *Science*, *339*(6122), 940-943.

Fatichi, S., D. Or, R. Walko, H. Vereecken, M. H. Young, T. A. Ghezzehei, T. Hengl, S. Kollet, N. Agam, and R. Avissar (2020), Soil structure is an important omission in Earth System Models, *Nat Commun*, *11*(1), 522.

Fatichi, S., et al. (2016), An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, *Journal of Hydrology*, *537*, 45-60.

Feng, D., K. Fang, and C. Shen (2020), Enhancing Streamflow Forecast and Extracting Insights Using Long□Short Term Memory Networks With Data Integration at Continental Scales, *Water Resources Research*, *56*(9).

Frame, J, Ulrich, P, Nearing, G, Gupta, H, Kratzert, F, (2022). On strictly enforced mass conservation constraints for modeling the rainfall-runoff process, EarthArXiv, doi: https://doi.org/10.31223/X5BH0P

Gao, S., Y. Huang, S. Zhang, J. Han, G. Wang, M. Zhang, and Q. Lin (2020), Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation, *Journal of Hydrology*, *589*.

Herrmann, F., N. Baghdadi, M. Blaschek, R. Deidda, R. Duttmann, I. La Jeunesse, H. Sellami, H. Vereecken, and F. Wendland (2016), Simulation of future groundwater recharge using a climate model ensemble and SAR-image based soil parameter distributions - A case study in an intensively-used Mediterranean catchment, *Sci Total Environ*, *543*(Pt B), 889-905.

Hoedt, P.-J., F. Kratzert, D. Klotz, C. Halmich, M. Holzleitner, G. S. Nearing, S. Hochreiter, and G. Klambauer (2021), MC-LSTM: Mass-Conserving LSTM, in *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Marina and Z. Tong, pp. 4275--4286, PMLR, Proceedings of Machine Learning Research.

Horton, R. E. (1933), The Rôle of infiltration in the hydrologic cycle, Eos Trans. AGU, 14(1), 446– 460, doi:10.1029/TR014i001p00446.

Jiang, S., Y. Zheng, and D. Solomatine (2020), Improving AI System Awareness of Geoscience Knowledge: Symbiotic

Integration of Physical Approaches and Deep Learning, *Geophys Res Lett*, *47*(13).

Jiang, S., Y. Zheng, C. Wang, and V. Babovic (2022), Uncovering Flooding Mechanisms Across the Contiguous United States Through Interpretive Deep Learning on Representative Catchments, *Water Resources Research*, *58*(1), e2021WR030185.

Jiang, S., Y. Zheng, V. Babovic, Y. Tian, and F. Han (2018), A computer vision-based approach to fusing spatiotemporal data for hydrological modeling, *Journal of Hydrology*, *567*, 25-40.

Jiang, S. J., V. Babovic, Y. Zheng, and J. Z. Xiong (2019), Advancing Opportunistic Sensing in Hydrology: A Novel Approach to Measuring Rainfall With Ordinary Surveillance Cameras, *WATER RESOURCES RESEARCH*, *55*(4), 3004-3027.

Jiang, Z.P., H.Y. Shi, S.N. Liu, Z.Q. Zhou, Y. Wang, and H.J. Cai (2021), Evolution characteristics of potential evapotranspiration over the Three-River Headwaters Region, *Hydrological Sciences Journal*, *66*(10), 1552-1566.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing (2019), Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, *Water Resources Research*, *55*(12), 11344-11354.

Kirkby, M.: Hydrograph modelling strategies, in: Processes in Human and Physical Geography, edited by: Peel, R., Chisholm, M., and Haggett, P., Heinemann, London, 69–90, 1975.

Leavesley, G.H., Lichty, R.W., Troutman, B.M., Saindon, L.G., 1983. Precipitation-Runoff-Modeling-System, User's Manual, Water Resource Investigations Report 83-4238, US Geological Survey

LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

Li, L., C. Bao, P. L. Sullivan, S. Brantley, Y. Shi, and C. Duffy (2017), Understanding watershed hydrogeochemistry: 2. Synchronized hydrological and geochemical processes drive stream chemostatic behavior, *Water Resources Research*, *53*(3), 2346-2367.

Liang, X., H. Zhan, and Y.-K. Zhang (2018), Aquifer Recharge Using a Vadose Zone Infiltration Well, *Water Resources Research*, *54*(11), 8847-8863.

Liu, S.N., and H.Y. Shi (2019), A recursive approach to long-term prediction of monthly precipitation using genetic

programming, *Water Resources Management*, *33*(3), 1103-1121.

Ma, Y., et al. (2022). On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence. https://doi.org/10.48550/arXiv.2207.04630.

Ma, Y., C. Montzka, B. Bayat, and S. Kollet (2021a), Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe, Hydrology and Earth System Sciences, 25(6), 3555-3575.

Ma, Y., C. Montzka, B. Bayat, and S. Kollet (2021b), An Indirect Approach Based on Long Short-Term Memory Networks to Estimate Groundwater Table Depth Anomalies Across Europe With an Application for Drought Analysis, *Frontiers in Water*, *3*.

Markstrom, S.L., Niswonger, R.G., Regan, R.S., Prudic, D.E., and Barlow, P.M., 2008, GSFLOW-Coupled Ground-water and Surface-water FLOW model based on the integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005): U.S. Geological Survey Techniques and Methods 6-D1, 240 p.

Mohan, C., A. W. Western, Y. Wei, and M. Saft (2018), Predicting groundwater recharge for varying land cover and climate conditions – a global meta-study, *Hydrology and Earth System Sciences*, *22*(5), 2689-2703.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I—A discussion of principles. Journal of Hydrology, 10(3), 282–290.

Newman, A. J., et al. (2015), Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrology and Earth System Sciences*, *19*(1), 209-223.

Niu, M., Horesh, L., & Chuang, I. (2019). Recurrent neural networks in the eye of differential equations. arXiv e-prints, arXiv:1904.12933. Retrieved from https://arxiv.org/abs/1904.12933

Ossai, I. C., A. Ahmed, A. Hassan, and F. S. Hamid (2020), Remediation of soil and water contaminated with petroleum hydrocarbon: A review, *ENVIRONMENTAL TECHNOLOGY & INNOVATION*, *17*.

Patil, S., and M. Stieglitz (2014), Modelling daily streamflow at ungauged catchments: what information is necessary?, *Hydrological Processes*, *28*(3), 1159-1169.

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–

707.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019), Deep learning and process understanding for data-driven Earth system science, *Nature*, *566*(7743), 195-204.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back□propagating errors. Nature, 323(6088), 533–536.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12:2493–2537.

Sadeghfam, S., A. Ehsanitabar, R. Khatibi, and R. Daneshfaraz (2018), Investigating 'risk' of groundwater drought occurrences by using reliability analysis, *Ecological Indicators*, *94*, 170-184.

Sahoo, S., T. A. Russo, J. Elliott, and I. Foster (2017), Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S, *Water Resources Research*, *53*(5), 3878-3895.

Su, D., H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao (2018), Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, Springer International Publishing, Cham.

Tan, L. S., Z. Zainuddin, and P. Ong (2018), Solving ordinary differential equations using neural networks, edited.

Tao, W., Q. Wang, L. Guo, H. Lin, X. Chen, Y. Sun, and S. Ning (2020), An enhanced rainfall–runoff model with coupled canopy interception, *Hydrological Processes*, *34*(8), 1837-1853.

Tian, Y., Y. Zheng, B. Wu, X. Wu, J. Liu, and C. Zheng (2015), Modeling surface water-groundwater interaction in arid and semi-arid regions with intensive agriculture, *Environ Modell Softw*, *63*, 170-184.Tran, H., E. Leonarduzzi, L. De la Fuente, R. B. Hull, V. Bansal, C. Chennault, P. Gentine, P. Melchior, L. E. Condon, and R. M. Maxwell (2021), Development of a Deep Learning Emulator for a Distributed Groundwater–Surface Water Model: ParFlow-ML, Water, 13(23).

Trinh, D. H., and T. F. M. Chui (2013), An empirical method for approximating canopy throughfall, *Hydrological Processes*, *27*(12), 1764-1772.

Viana, F. A. C., R. G. Nascimento, A. Dourado, and Y. A. Yucesan (2021), Estimating model inadequacy in ordinary differential equations with physics-informed neural networks, Computers & Structures, 245.

Wang, N., Zhang, D., Chang, H., & Li, H. (2020). Deep learning of subsurface flow via theory-guided neural network. Journal of Hydrology, 584, 124700.

Wang, N., H. Chang, and D. Zhang (2021), Deep□Learning□Based Inverse Modeling Approaches: A Subsurface Flow Example, *Journal of Geophysical Research: Solid Earth*, *126*(2).

Yoon, H., S.-C. Jun, Y. Hyun, G.-O. Bae, and K.-K. Lee (2011), A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, Journal of Hydrology, 396(1-2), 128-138.

Zhang, J., X. Chen, A. Khan, Y.-k. Zhang, X. Kuang, X. Liang, M. L. Taccari, and J. Nuttall (2021), Daily runoff forecasting by deep recursive neural network, *Journal of Hydrology*, *596*.

Zhao, W. L., P. Gentine, M. Reichstein, Y. Zhang, S. Zhou, Y. Q. Wen, C. J. Lin, X. Li, and G. Y. Qiu (2019), Physics-Constrained Machine Learning of Evapotranspiration, *Geophys Res Lett*, *46*(24), 14496-14507.

Zhou, Z., H. Shi, Q. Fu, T. Li, T. Y. Gan, and S. Liu (2020), Assessing spatiotemporal characteristics of drought and its effects on climate-induced yield of maize in Northeast China, *Journal of Hydrology*, *588*.

Zipper, S. C., T. Dallemagne, T. Gleeson, T. C. Boerman, and A. Hartmann (2018), Groundwater Pumping Impacts on Real Stream Networks: Testing the Performance of Simple Management Tools, *Water Resources Research*, *54*(8), 5471-5486.

Author contributions

Conceptualization: Haiyun Shi

Methodology: Hejiang Cai, Haiyun Shi

Formal analysis: Hejiang Cai

Validation: Hejiang Cai, Zhaoqiang Zhou, Shijie Jiang

Software: Hejiang Cai, Zhaoqiang Zhou

Supervision: Haiyun Shi, Suning Liu, Vladan Babovic

Writing - original draft: Hejiang Cai

Writing - review & editing: Haiyun Shi, Suning Liu, Vladan Babovic, Shijie Jiang

Funding Acquisition: Haiyun Shi