Environmental Science and Ecotechnology 17 (2024) 100298

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Deep-reinforcement-learning-based water diversion strategy

Qingsong Jiang [a], Jincheng Li [a], Yanxin Sun [a], Jilin Huang [a], Rui Zou [b], Wenjing Ma [b], Huaicheng Guo [a], Zhiyun Wang [c], Yong Liu [a, *]

[a] State Environmental Protection Key Laboratory of All Materials Flux in River Ecosystems, College of Environmental Sciences and Engineering, Peking University, Beijing, 100871, PR China
[b] Rays Computational Intelligence Lab, Beijing Inteliway Environmental Ltd., Beijing, 100085, PR China
[c] Yunnan Key Laboratory of Pollution Process and Management of Plateau Lake-Watershed, Yunnan Research Academy of Eco-environmental Sciences, Kunming, 650034, PR China

ARTICLE INFO

Article history:
Received 20 December 2022
Received in revised form
23 June 2023
Accepted 5 July 2023

Keywords:
Dynamic water diversion optimization
Deep reinforcement learning
Process-based model
Explainable decision-making
Parameter uncertainty

ABSTRACT

Water diversion is a common strategy to enhance water quality in eutrophic lakes by increasing available water resources and accelerating nutrient circulation. Its effectiveness depends on changes in the source water and lake conditions. However, the challenge of optimizing water diversion remains because it is difficult to simultaneously improve lake water quality and minimize the amount of diverted water. Here, we propose a new approach called dynamic water diversion optimization (DWDO), which combines a comprehensive water quality model with a deep reinforcement learning algorithm. We applied DWDO to a region of Lake Dianchi, the largest eutrophic freshwater lake in China and validated it. Our results demonstrate that DWDO significantly reduced total nitrogen and total phosphorus concentrations in the lake by 7% and 6%, respectively, compared to previous operations. Additionally, annual water diversion decreased by an impressive 75%. Through interpretable machine learning, we identified the impact of meteorological indicators and the water quality of both the source water and the lake on optimal water diversion. We found that a single input variable could either increase or decrease water diversion, depending on its specific value, while multiple factors collectively influenced real-time adjustment of water diversion. Moreover, using well-designed hyperparameters, DWDO proved robust under different uncertainties in model parameters. The training time of the model is theoretically shorter than traditional simulation-optimization algorithms, highlighting its potential to support more effective decision-making in water quality management.

1. Introduction

Global lake eutrophication and algal blooms are strongly linked to excess nitrogen and phosphorus [1]. Reducing external nutrient inputs and lowering lake nitrogen and phosphorus concentrations are long-term goals in order to mitigate the risk of algal blooms. This is accomplished through widely adopted countermeasures, such as pollution source interception, robust sewage treatment, dredging, wetland restoration, and water diversion [2]. However, under the compounding impacts of increasingly intense anthropogenic activities and extreme weather events, the state of the lake water quality has become increasingly unstable [3]. The environmental and ecological effects of these countermeasures are declining, and water quality improvement and algal bloom control are met with significantly greater challenges [4]. Therefore, there is an urgent need to improve the effectiveness of existing measures to improve lake water quality.

Among various solutions, inter-basin water diversion projects have been increasingly implemented in recent years, aiming to increase the available water resources, accelerate water circulation, and improve lake water quality. Some notable cases that implemented these solutions are the South-North Water Diversion Project [5]. The Water Diversion Project from the Yangtze River to Lake Tai [6] and the Niulan River—Dianchi Water Diversion Project [7,8]. Statistics have shown that the average annual water diversion for ecological and environmental goals has exceeded 30 billion m^3 in China. Although the external water resource accelerates circulation, there are significant debates surrounding its utility since it

* Corresponding author.
E-mail address: yongliu@pku.edu.cn (Y. Liu).

https://doi.org/10.1016/j.ese.2023.100298

substantially increases the nitrogen and phosphorus input loads as well as the pressure from removing these nutrients within the lake. As a result, the water quality improvement was much lower than the expected goals [9,10]. From the perspective of restoring water quality, robust water diversion decisions need to incorporate three aspects: (a) the demand for a lake that supplies external water resources (e.g., the higher the concentration of nitrogen and phosphorus in the lake, the larger the diversion); (b) the ability of external water resources to improve the water quality of the lakes (e.g., the lower the nitrogen and phosphorus concentration of the external water, the larger the improvement); and (c) the perturbation of other factors (e.g., the discharge of the tributaries and meteorological factors) [11–14]. Once the external and internal factors change, according to the optimality theory, water diversion decisions must be adjusted. Therefore, it becomes a dynamic optimization problem, where the current water diversion affects the concentration of the lake during the following period and the next diversion decision. The water diversion projects, however, have yet to address adequately the issue of efficient and timely decision making.

Solving dynamic optimization problems cannot rely solely on traditional algorithms, as the objective function, constraints, and Pareto front surface may change over time [15]. Particle swarm optimization (PSO) and genetic algorithms (GA) are more suitable algorithms for solving static (non-dynamic) optimization problems [16], while their variants have difficulty in balancing all-period optimality as well as dynamic computational effort issues [17]. Reinforcement learning (RL) was created for dynamic optimization as a machine learning algorithm branch. Through trial and error, RL optimizes the decision-making strategies in order to achieve the highest cumulative reward [18] and possesses the advantages of exploration-exploitation, diverse training methods, and flexible dependence on the *Environment* (i.e., the world where agents interact and are observed). Unlike supervised and unsupervised learning, RL does not require any labeled data; however, it interacts with the *Environment* and obtains data (states and rewards) that update the decision-making strategies. Early RLs use matrices to store reward information, which is only suitable for discrete state and action spaces [19]. Deep learning (DL), skillful in approximating any function, is coupled with RL and produces modern deep reinforcement learning (DRL) to solve issues in continuous spaces. With powerful performance, DRL is widely used in robotics, autonomous driving, and games [20–23].

RL has been used in water resource scheduling with better performance than traditional dynamic programming methods [24–27]. In water quality management, the RL was coupled with the *Environment* of the process-based models. It is aimed at optimal decisions in various cases, such as: (a) controlling the discharge of reservoirs to ensure downstream water quality [28]; (b) keeping dissolved oxygen and nitrate concentrations within the wastewater treatment process in order to reduce energy consumption and effluent pollutants [29–31]; and (c) controlling the opening degree of the valve to reduce rain-induced flooding and suspended matter within the drainage systems [32,33]. As for restoring lake eutrophication, complex nonlinear water quality response shows an increasing demand for RL that optimizes the project operations. Within the lakes, nonlinearity arises due to the variability of external boundary conditions and the heterogeneity of the internal processes [34]; however, DRL can handle the complexity. Even in an incompletely observable *Environment*, DRL can obtain an optimal strategy based on interactions with the *Environment* and Bellman's Principle of Optimality. The *Environment* mentioned here is a real system whose state and reward are monitored by sensors or a virtual system consisting of process-based models that significantly reduce the DRL training time.

However, DRL is still not adequately applied to lake water quality restoration efforts compared to the potential demands. The possible reasons are: (a) although DRL training does not require numerous labeled data, an available *Environment* only requires data from monitoring or surrogate models. Additionally, constructing complex surrogate models is not easy, and the hyperparameters of a model-based *Environment* are difficult to design, such as the reward function [35]; (b) computationally, the response of the *Environment* to the actions generated by the DRL is a time-consuming process, difficult to converge, and is the primary reason why many current studies use simple models; and (c) it is well known that interpretability is a typical machine learning algorithm problem (i.e., DRL cannot explain how to make decisions as well as identify the specific inputs that determine the prediction of the decisions). Nevertheless, water diversion decisions are normally driven not only by the water quality of the transferred water and the state of the lakes but also by the operating costs and future meteorological conditions. Therefore, it is essential to identify important inputs needed for deeper management. However, these shortcomings in interpreting the “black box” have reduced the reliability of DRL and hindered its application.

This study aims to solve the aforementioned problems by proposing a dynamic water diversion optimization method (DWDO) for water diversion projects within eutrophic lakes. To maximize the water quality improvement of water diversion and reduce operation costs, DWDO generated the dynamic optimal amount of water diversion (AWD) under changing external conditions. A three-dimensional hydrodynamic-water quality-algae model based on the Environmental Fluid Dynamics Code (EFDC) and a novel Deep Deterministic Policy Gradient (DDPG) coupling DL with RL were combined in DWDO. The water quality response simulation and the AWD acted as the bridge connecting the *Environment* and DRL. We tested DWDO in the case of Lake Dianchi, which is one of the three most eutrophic large lakes in China. Furthermore, we gained insight into the decision-making mechanism within DDPG using an explainable approach. This approach identified the key driving factors and their contribution to diversion decisions, providing end-to-end support for eutrophication control.

2. Methods

The proposed DWDO in this study consisted of (a) the *Environment* and (b) DDPG (Fig. 1). A complex process-based water quality model was adopted as a surrogate for the *Environment*. To accurately quantify the nitrogen, phosphorus, and algae variations in the eutrophic lake, we simulated the spatiotemporal changes of the hydrodynamics, temperature, sediment, and algae, as well as water quality. This requires a wealth of data for EFDC (e.g., meteorological data, water quantity, and inlet load data), water level observation, and the lake's water quality. This data is used as the initial and boundary conditions or to calibrate the parameters of the model. Additionally, a certain amount of data is necessary for complex problems, such as dynamic water diversion. Based on the EFDC simulation, we were able to calculate the boundary fluxes (e.g., inflow flux, outflow flux, benthic flux, atmospheric deposition, and denitrification) as well as the internal fluxes of the lake (e.g., algal uptake, mineralization, and hydrolysis) at the present step and the meteorological condition statistics and water diversion water quality during the next step. They constituted the state space, potentially impacting the AWD generated by DDPG. DDPG learned strategies under these high dimensional continuous state spaces and generated AWD in real time. Moreover, by interacting with the *Environment*, DDPG collected enough data to update the neural network's weights until the convergence of the objective function.

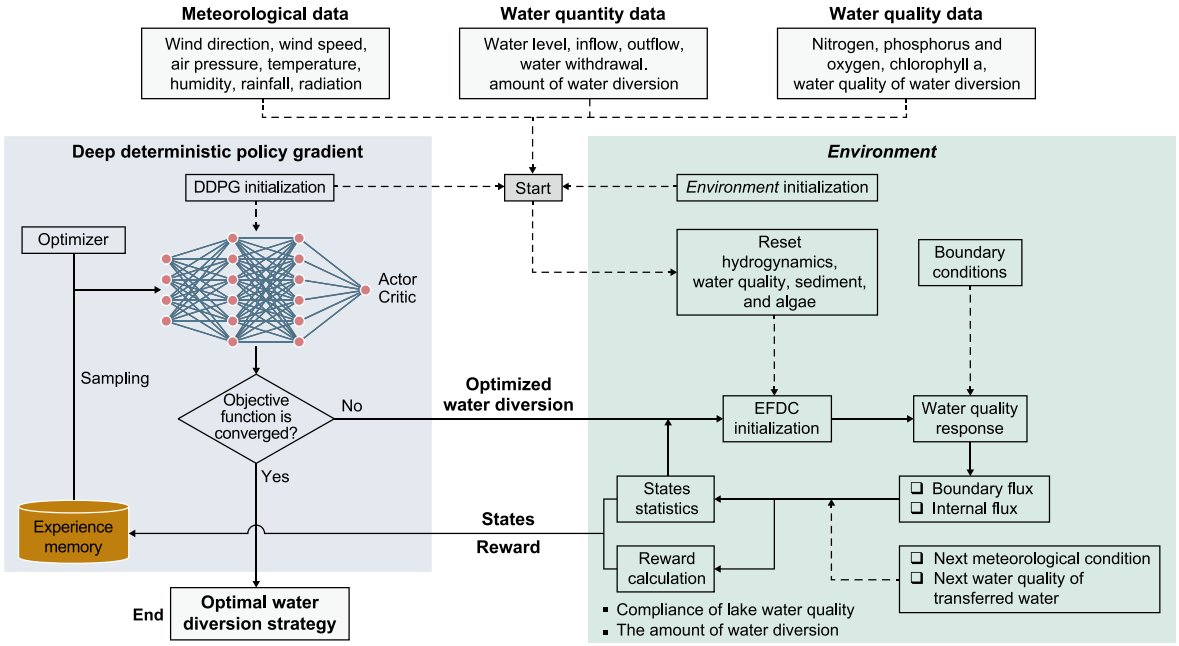

Fig. 1. The DWDO framework of this study. The dashed arrows represented the data input or the initialization of the *Environment* and DDPG. The solid arrows represented the interaction processes between and within the *Environment* and DDPG.

2.1. Three-dimensional hydrodynamic-water quality-algae simulation

To meet the requirements of high spatial and temporal resolution, we used Intelligent Watershed Integrated Decision-making—Lake & Reservoir (IWIND-LR), an upgraded version of EFDC, as the modeling platform. IWIND-LR has inherited the advantages of EFDC in simulating the hydrodynamics and improved the water quality and sediment modules [36]. The basic form of the governing equation is as follows:

$$\frac{\partial C}{\partial t}+\frac{\partial (uC)}{\partial x}+\frac{\partial (vC)}{\partial y}+\frac{\partial (wC)}{\partial z}-\frac{\partial}{\partial x}\left(K_x\frac{\partial C}{\partial x}\right)-\frac{\partial}{\partial y}\left(K_y\frac{\partial C}{\partial y}\right)-\frac{\partial}{\partial z}\left(K_z\frac{\partial C}{\partial z}\right)=S+R+\mathrm{Q} \quad (1)$$

where; C is the concentration of the simulated variable; t is time; u, v, and w are velocity components in the x, y, and z directions, respectively; K_x, K_y, and K_z are the turbulent diffusivities in the x, y, and z directions, respectively; S is the net value of deposition and release, R is the flux of the internal biochemical processes, and Q is the net value of the inflow and outflow. Variables in IWIND-LR include carbon, nitrogen, and phosphorus present in refractory particulate organic, labile particulate organic, dissolved organic, and inorganic forms, as well as algal biomass and dissolved oxygen. All of these variables were modeled in this study. In addition, it was necessary to simulate fluxes within the sediment and their interactions with the overlying water due to the significant contribution of sediments in eutrophic lakes [37].

In this study, Lake Caohai in the northern region of Lake Dianchi was used as the study area. Lake Caohai, with an area of 10.8 km^2 and a watershed area of 153.4 km^2, is surrounded by an urban area and is subject to large exogenous nitrogen and phosphorus loads (Fig. 2). Overlaying with the limited water resources in the basin, the local government had to implement an expensive water diversion project (i.e., the Niulan River—Dianchi Water Diversion project (NDWD)) [8]. The high anthropogenic activity intensity and the warm plateau climate made the benefits of water diversion nonlinear and uncertain.

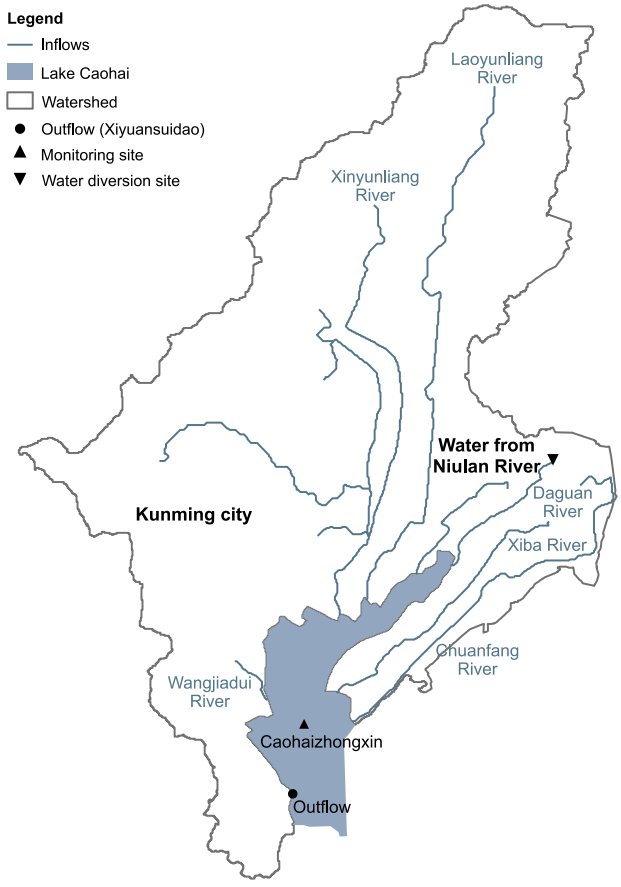

Fig. 2. Study area of the monitoring site and spatial watershed information.

We collected data on the bottom elevation, hourly meteorological conditions, daily water quantity, water level, the monthly water quality of the inlets and the lake, and the daily and monthly water quality of water diversion (Text S1). After discretizing the water surface, there were 319 horizontal grids with a total area of 7.12 km^2. In the vertical direction, it was divided into two layers with an average depth of 1.25 m. Since the NDWD was constructed in 2016, the modeling time of this study was set as 2017−2020 with a computational time step of 100 s.

2.2. Architecture of the Environment and hyperparameter configuration

Following the design of the OpenAI Gym, the architecture of the *Environment* contained the necessary functions of *reset*, *initialize*, and *step*. When a training episode finished, the function *reset* set all states of the *Environment*, including time, hydrodynamics, water quality, sediment, and algae, as their default values. The *Environment* then ran the function *initialize* to set the start time, episode duration, state and action spaces, action duration, water quality targets, and their weights (Table 1).

Within one episode, the function *step* was responsible for the interaction with DDPG. This function contained four substeps:

(1) Model initialization: When the previous step ended, a total of 76 hydrodynamic, 27 water quality, and 18 sediment variables, as well as the temperature values of the entire area (319 grids), were used as initial values for the next simulation. This approach was consistent with the fact that the state of the lake was consecutive at the moment of decision adjustment.
(2) Model running: Similarly, the time series of the water quantity, water quality, and water temperature of each inlet and nine meteorological factors were screened out from the collected data and used as the models' boundary conditions during the next step. Under AWD by DDPG, the model ran continuously.
(3) States statistics: The variables simulated by the model cannot be directly fed to the DDPG due to the extremely high dimension of hundreds of variables in hundreds of grids. To reduce the dimension, we calculated all fluxes occurring at or within the lake, as well as the total mass of nitrogen and phosphorus in the overlying water (Text S1). Due to the different flux and mass magnitudes, they were normalized (Table S1) [38]. In addition, we added the maximum, minimum, and average values of the meteorological factors of the next step as well as the future total nitrogen (TN) and total phosphorus (TP) of the water diversion (TN_{wd} and TP_{wd}). In summary, the state space consisted of 67 dimensions.
(4) Reward calculation: The reward was the core component of DRL. From the management perspective, we first brought the compliance rate of TN and TP into the reward function and named it the water quality reward. This function was 0 when the TN and TP of the lake met the standards; otherwise, a negative reward was feedback (equation (3)). If water diversion failed to mitigate lake water quality, we then expected that DDPG would reduce AWD in time. We next added a water diversion reward into the total reward; therefore, the total reward is calculated as follows:

$$R = k_{wq}R_{wq} + k_d R_d \tag{2}$$

$$R_{wq} = -\frac{1}{S}\sum_{s=1}^{S}\max\left(\frac{\frac{1}{T}\sum_{t}^{T} wq_{s,t}}{wq_{s,target}} - 1, 0\right) \tag{3}$$

$$R_d = -\frac{Q_d}{Q_{max}} \tag{4}$$

where R_{wq} and R_d denote the water quality and diversion rewards, respectively; k_{wq} and k_d are their weights, respectively; and $wq_{s,t}$ is the concentration of variable s at time t (mg L^{-1}); $wq_{s,target}$ is the water quality target (mg L^{-1}); Q_d is the AWD (m^3 d^{-1}), and Q_{max} is the maximum AWD (m^3 d^{-1}).

The $wq_{s,target}$ and weights (k_{wq} and k_d) are the key configurations of the reward function. A too low $wq_{s,target}$ will lead to R_{wq} being less than 0 for an extended period of time, compelling DDPG to continuously increase the AWD in order to improve R_{wq}. Too high a $wq_{s,target}$ will lead to R_{wq} often being equal to 0, where water diversion can easily achieve the proposed goals. The reward weights of k_{wq} and k_d were set to balance R_{wq} and R_d. The ratio of k_{wq}/ k_d determined whether the DDPG paid more attention to water quality or water diversion. The k_{wq} should be much greater than k_d, since water diversion aims to improve water quality.

2.3. DDPG and its hyperparameters configuration

DDPG was used to solve the dynamic decision problems in continuous state and action spaces. It conducted a deterministic selection to pick up the action with the highest probability. In terms of the algorithms, DDPG was similar to the architecture of the Actor-Critic and had a total of four sub-networks: (a) Actor-generated action a_0 based on the current state s_0; (b) Critic computed the corresponding objective function $Q(s_0,a_0)$; (c) Target Actor generated the action a_1 according to the next step state s_1; and (d) Target Critic computed the corresponding objective

Table 1
Initialization of the *Environment* in DRL.

Variable	Value	Unit	Description
Start time	1	day	The time when the episode started, such as the first day of the year.
Episode duration	365	day	The maximum days when an episode ended. In the end, the start time and initial state need to be reset.
Maximum AWD	15	m^3 s^{-1}	The maximum amount of water diversion.
Action space	[0,15]	-	The range of AWD to be transferred is similar to the Gym's spaces.
Action duration	7	day	the number of days the model ran in one step, where AWD and the water quality of water diversion were deemed constant.
State space	Norm (μ, δ^2)	-	the normalized space of variables of the observable *Environment*
Water quality targets	TN = 2 TP = 0.05	mg L^{-1}	The target concentrations of TN and TP of the lake through water diversion.
Weights of water quality targets	k_{wq} = 99 k_d = 1	-	The coefficients of water quality reward and water diversion reward when calculating the total reward.

function $Q'(s_1, a_1)$. The Actor, the Critic, and the Target Critic have the same network architecture as the Target Actor. The objective function of the Actor is:

$$Q(s_m, \theta) = E\left(\sum_{i=1}^{M} \gamma^{i-m} R_i \middle| s_m, \theta \right) \tag{5}$$

$$\text{Max } J(\theta) = -\frac{1}{M} \sum_{m=1}^{M} Q(s_m, \theta) \tag{6}$$

And Critic's objective function is:

$$\text{Min } J(w) = \frac{1}{M} \sum_{m=1}^{M} (R + \gamma Q'(s_{m+1}, a_{m+1}, w') - Q(s_m, a_m, w))^2 \tag{7}$$

where, θ, w, θ', and w' were the weights of the Actor, the Critic, the Target Actor, and the Target Critic, respectively, $Q(s_m, \theta)$ was the objective function calculated by the expectation value of the cumulative sum of reward R, γ is the discount factor, and M is the batch of samples used to update the network's weights.

The DDPG training process is as follows:

(1) Initializing randomly θ, w, θ', and w' being the same as θ and w, respectively;
(2) Based on the current s_0, the Actor generated a_0. The *Environment* responded and feeds back the s_1 and reward. This information was stored in the experience replay database Memory;
(3) Sampling from Memory, the Critic calculated the Q values based on s_m, a_m, R_m, and s_{m+1}. They updated θ and w via an optimizer;
(4) Updating θ' and w' using the soft update method;
(5) Assigning s_1 as s_0, and repeating from (2) to (4) until DDPG converges.

The Actor, the Critic, and their target networks comprised fully-connected neural networks (FNNs), whose powerful nonlinear representation was a huge advantage. However, the architecture of DDPG and FNNs contained important hyperparameters: (a) discount factor, batch size, warm start steps, training interval, experience storage frequency, and soft update factor in DDPG, and (b) network architecture, weights initialization, activation function, optimizer, and learning rate in FNN. All hyperparameters must be adjusted to adapt to the needs of the complex water diversion problems.

In DDPG, the discount factor γ had a significant impact. Moreover, γ being closer to 1 represented more attention to the rewards of future lake water quality under this AWD; in this study, it was 0.9. The batch size represented the number of samples used for weight updating. The larger the batch size, the more robustly the Actor and Critic weights were updated. However, a larger batch size required more training data. This study was set at 32. The warm start step was set at 64, which is used to collect enough training samples for a stable beginning before an update. In addition, the DDPG soft update factor represented the update speed of the Target Actor and the Target Critic. The larger the value, the faster the weights update; however, it may lead to the convergence difficulty of the algorithm; therefore, it was set to 0.001.

In FNN, the architecture of the neural network is related to its ability to fit complex relationships, convergence, and training time [39]. In this study, both the Actor and the Critic had two hidden layers (Table 2). The Actor used ReLu as the activation function of the hidden layer and then used Gaussian noise to enhance the noise capability. Its output layer was a tanh function with a range of (−1, 1). The Critic is similar to the Actor, but its output layer is a linear function. The number of weights of the Actor and the Critic is approximately 50,000, which characterizes the high nonlinear fitting ability. These weights were initialized randomly. The optimizer in the training process was the efficient Adam optimizer [40], whose learning rate was set to 0.001.

2.4. Explanation of water diversion strategies

Since machine learning algorithms cannot explain the decisions independently, we applied an independent post-processing tool, SHAP (SHapley Additive exPlanations). SHAP is a game theory method for explaining the output-input relationship of any machine learning algorithm [41], surpassing other additive feature attribution methods. It approximated the original model $f(x)$ (a black box) and simplified it into an additive explainable model. With SHAP, the $f(x)$ was attributed to the addition of the marginal contribution of each input variable as follows:

$$g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i' \tag{8}$$

where, g is the explanation model; $z_i' \in \{0, 1\}$, $z_i' = 1$ if the corresponding input variable exists; M is the number of input variables, and the attribution value $\varphi_i \in R$. If the original model was a simple linear regression method, the SHAP value could be interpreted vividly as the input value multiplied by its coefficient. If the original model was complex, the SHAP value of input A is:

$$SHAP(A) = \frac{1}{S} \sum_{A}^{S} f(A, B) - f(B) \tag{9}$$

where, B is the variable set that excluded A, and its predicted value was $f(B)$; $f(A, B)$ is the prediction when A is added into the model, and $f(\text{A}, \text{B}) - f(B)$ is the contribution of A. However, this contribution might vary with the value of set B, so it is necessary to calculate the average contribution of A under different sets of B by sampling. As a result, the SHAP value quantified not only the global contribution of the variables to the decision but also the local contribution of the inputs to the prediction in each sample.

Table 2
The network architecture of the Actor and the Critic in DDPG.

Component of Architecture	Actor	Critic
Input layer	States	States, action
Dimension of the input layer	67	68
Size of first hidden layer	256	256
The activation function of the first layer	ReLu	ReLu
Number of weights in the first layer	17408	17664
Output processing of the first layer	Gaussian noise	-
Size of second hidden layer	128	128
The activation function of the second layer	ReLu	ReLu
Number of weights in the second layer	32896	32896
Output processing of the second layer	Gaussian noise	-
Output layer	Action	Q-value
Dimension of the output layer	1	1
The activation function of the output layer	Tanh	Linear
Number of weights in the output layer	129	129
Total number of weights	50443	50689

3. Results and discussion

3.1. Evaluation of the water quality improvement for NDWD

The calibrated water quality model showed that the Nash-Sutcliffe Efficiency coefficient (NSE) of the water level, surface water temperature, and evapotranspiration were 0.27, 0.80, and 0.58, respectively. The model reproduced the hydrodynamical trends and the biases were acceptable (Fig. 3a–c). The water quality model parameters were set within the rational range from the literature (Table S2). The calibration showed that the simulated TN, TP, and chlorophyll *a* (Chla) values fit well with the overall observed trend with an NSE of 0.08, 0.09, and 0.03, respectively. Therefore, considering the uncertainty during extreme observation resulting from the considerable divergence between the observed data at the specified time and the data collected at the time points prior to and subsequent to the observation (Fig. 3d and e).

The historical NDWD has an average AWD of 8.1 $m^3\ s^{-1}$ and a maximum AWD of 15 $m^3\ s^{-1}$. However, the water from the Niulan River can neither lead to a continuous decrease of TN and TP in the lake nor improve the water quality of the whole lake throughout the year (Fig. 4). On the basis of the operation data and scenario simulations generated by EFDC the AWD was reduced when the TN_{wd} or TP_{wd} was high (e.g., in October 2017); conversely, AWD increased (e.g., in spring 2018). The lake's water quality showed a high consistency with the Niulan River. Its concentrations did not exceed the TN_{wd} and TP_{wd}, indicating that the internal circulation of the lake enabled water quality concentrations to further decrease. The improvement of TP concentration was lower than that of TN, in the past operation of NDWD. This was due to the AWD not adjusting in time according to the TN_{wd} and TP_{wd}. For example, in June 2018, both TN_{wd} and TP_{wd} were high at 6.0 and 0.15 mg L^{-1}, respectively; however, the NDWD still transferred water to Lake Caohai at a rate of greater than 10 $m^3\ s^{-1}$, increasing TN and TP of the lake that grew higher than the concentrations in the non-diversion scenario. Therefore, the NDWD should be decreased when TN_{wd} and TP_{wd} was significantly greater to improve water quality and save operational costs.

The NDWD provided an average diversion of 528 million m^3 per year. With the average local water price at CNY 2 m^{-3}, the annual cost was more than one billion yuan. Unreasonable diversion decisions led to higher operating costs. We expected that less AWD could achieve the same or a better effect on water quality; therefore, greatly reducing the water diversion cost. A more accurate decision could be made using the multi-level valve allocation (Text S2).

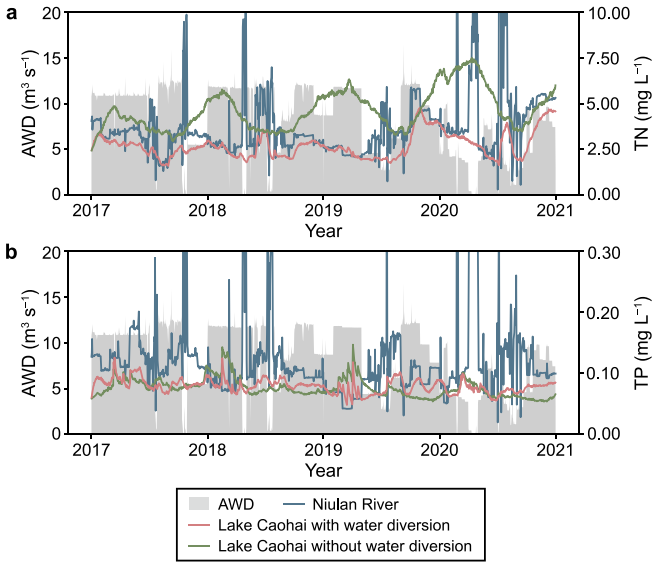

Fig. 4. The comparison of TN (**a**) and TP (**b**) of the Niulan River, and the Lake Caohai with and without water diversion.

3.2. Training strategies of DWDO

In this study, DWDO was trained for 300 episodes (i.e., 300 years of simulation) until the objective function converged (Fig. 5a and b). Since each episode contained 52 steps (spending 0.4 h), 15,000 training samples were collected for the DDPG. The training task was performed on a personal computer, with the most time-consuming process being the response simulation of the *Environment*. For complex process-based EFDC, each time step required tens of minutes to compute numerical solutions for thousands of grids and variables. The DWDO training process consisted of three stages (Fig. 5c): (a) Maximum AWD strategy. At this stage, DDPG adopted

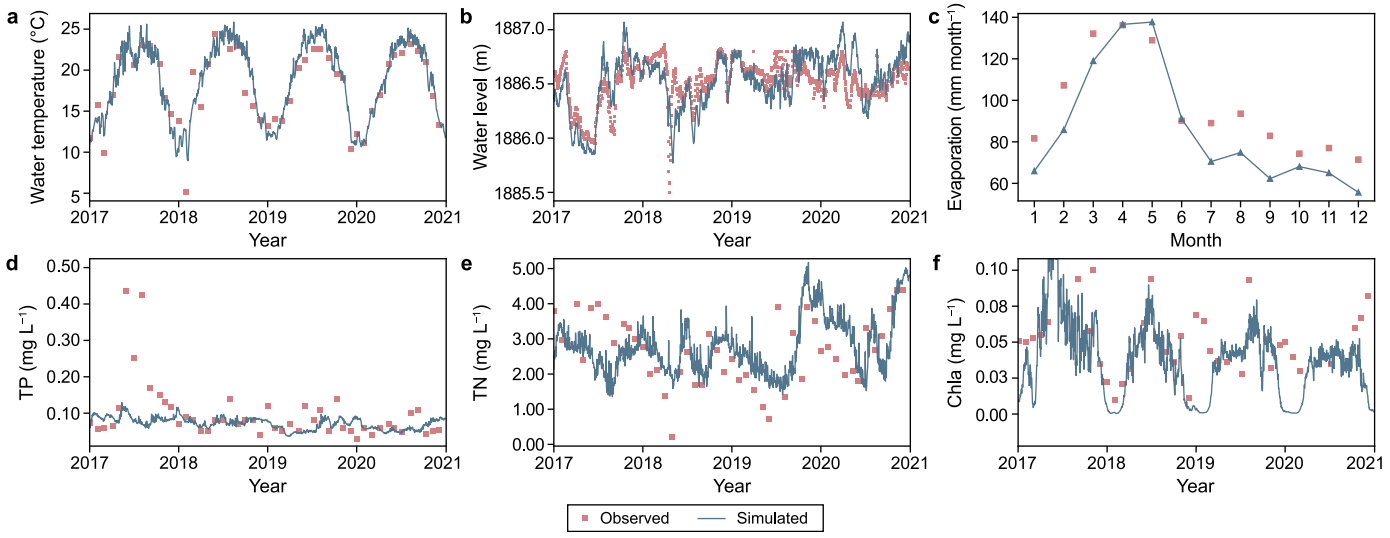

Fig. 3. Calibrations of the process-based model and the trends of surface water temperature (**a**), water level (**b**), evaporation (**c**), TP (**d**), TN (**e**), and Chla (**f**).

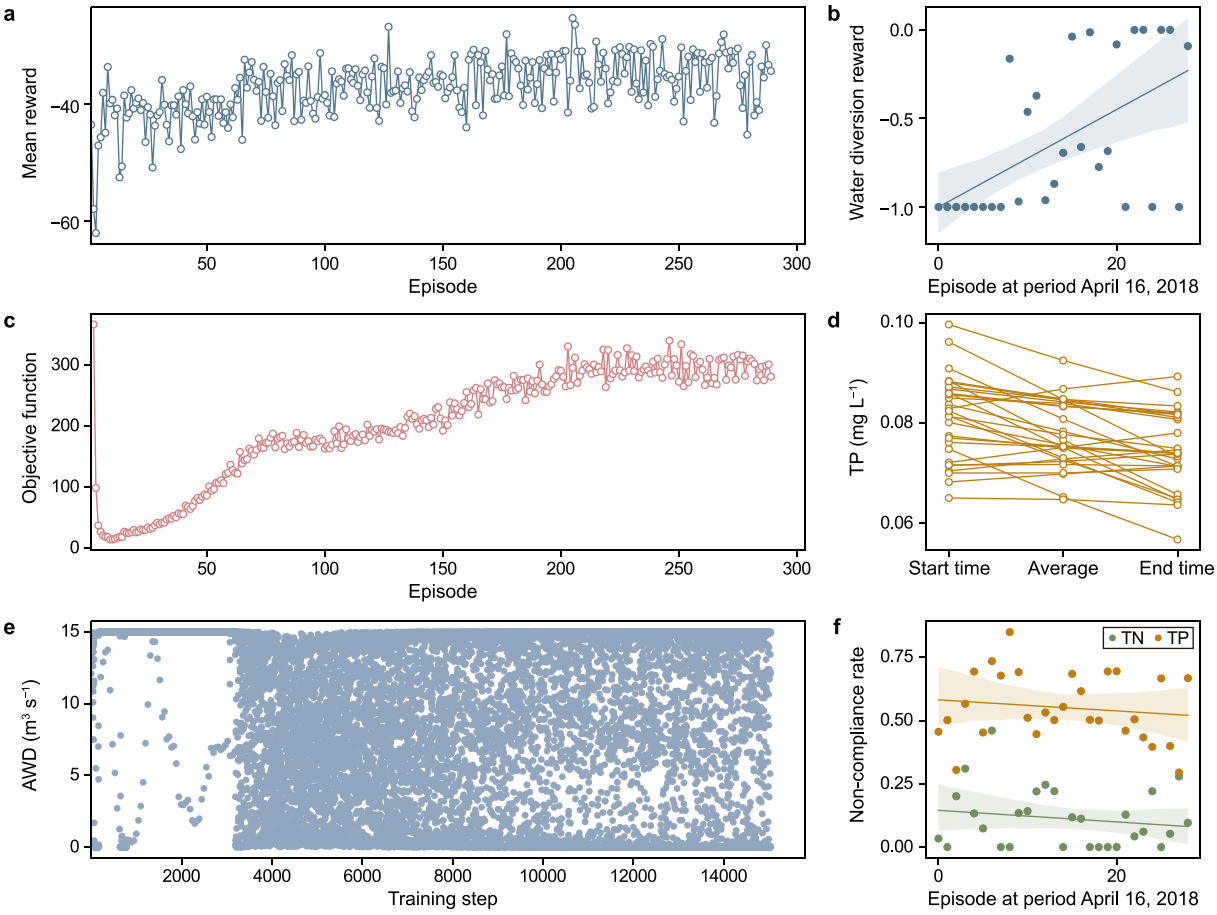

Fig. 5. Training of the DWDO and its change of rewards on water quality and water diversion. **a**, Mean reward of each episode. Its fluctuation was caused by randomness in the training process. **b**, Water diversion reward on April 16, 2018. **c**, The objective function of the Actor in DDPG. **d**, The decrease of TP at the end of one step. **e**, The amount of water diversion (AWD), step by step. **f**, The decrease of the non-compliance rate of TN and TP of Lake Caohai on April 16, 2018.

an aggressive strategy of increasing water diversion to improve water quality and rewards; however, it was found that increasing AWD at high TN_{wd} and/or TP_{wd} did not improve lake water quality; (b) Random strategy. DDPG shifted to a stochastic strategy, i.e., it generated AWD from a uniform distribution. This strategy was still not efficient enough to improve water quality; (c) Combining the experience of the first two stages, the strategy of the third stage was biased towards decisions with larger water quality and diversion rewards (Fig. 5d). After convergence, the DDPG found a dynamic balance between the water diversion and the water quality rewards. In summary, the non-compliance rate of the TN and TP of the lake did not decrease (Fig. 5e and f). Consequently, the rise of rewards showed that water diversion improved lake water quality. Additionally, the water diversion reward increased, revealing that the dual optimization of the TN and TP of the lake and water diversion was successful.

3.3. Verification and explanation of decisions

To demonstrate the DWDO-induced water quality improvement, we compared the TN and TP changes in Lake Caohai between AWDs by DWDO and the observed AWD from 2017 to 2020. We set seven days as the step and obtained a total of 208 dynamical decision steps. The optimal diversion strategy produced a better lake water quality than that under the past diversion rules (Fig. 6a and b). The results showed that DWDO decided to reduce water or stop diversion when the TN_{wd} or TP_{wd} was high (e.g., in June 2018, October 2019, and July and November 2020). Conversely, it decided to transfer more water or even keep the maximum AWD (e.g., in January 2019). The DWDO strategy successfully led to a significant decrease in the TN and TP concentrations. Compared with the past diversion rules, the TN and TP of Lake Caohai decreased by 7% and 6%, respectively, from 2017 to 2020. Conversely, the total AWD by DWDO decreased dramatically to an average value of approximately 60 million m^3 per year. It was 75% lower than the observed AWD, and the total reward rose (Fig. 6c and d). Technically, it succeeded in achieving the dual goals of water quality improvement and cost saving.

We gained insight into the attribution of all states of the water diversion decisions. Under the optimal strategy, the largest contributor to the decision most often was the TN_{wd} and TP_{wd} (Fig. 7a and b). Their average SHAP values reached −0.45 and −0.82, respectively. Negative SHAP values indicated that the high TN_{wd} and TP_{wd} forced DWDO to turn down the AWD, rather than to increase it at low TN_{wd} and TP_{wd} (red points with SHAP >0). In contrast, the lake's dissolved organic nitrogen (DON) mass asked for an increase in AWD with a mean SHAP value of 0.13. Furthermore, meteorological factors had a lower contribution to the decisions than TN_{wd} and TP_{wd}, as well as the masses and the fluxes of nutrients. However, they still caused a significant increase or decrease in AWD. Wind speed had higher SHAP values than other meteorological factors, which might be related to the fact that the

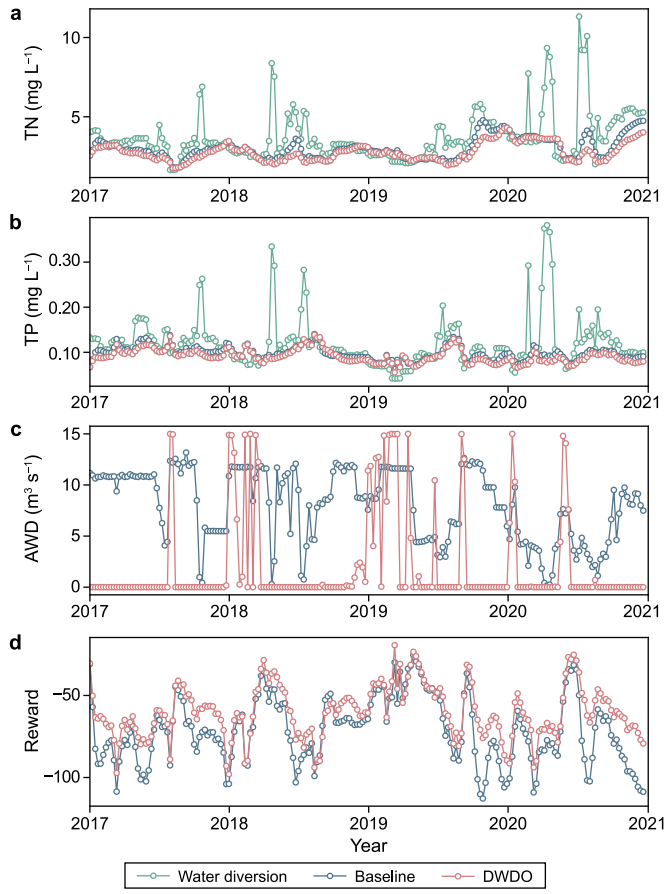

Fig. 6. Verification of DWDO's performance compared to the past diversion rule. **a**, TN; **b**, TP; **c**, The amount of water diversion (AWD); **d**, The total reward.

particulate nitrogen and phosphorus of eutrophic shallow lake sediments are resuspended by wind-wave action and transported via lake circulation [42].

The decisions could be driven by the coupling effects of multiple factors, especially from December to June (Fig. 7b). Among these significant factors, the sum of the factors with a positive SHAP revealed a consistent trend with the negative factors (Fig. 7c). However, the sum of all factors primarily existed in the winter and spring, which determined the specific AWD. If the positive SHAP increased or the negative SHAP decreased, the AWD generated by DWDO would increase, and vice versa.

Furthermore, if a factor value is too large or too small, it might increase or decrease AWD (Fig. 7d–h and Fig. S1). For example, the higher the PO_4 mass of the lake, the higher the AWD calculated by DWDO due to the increasing demand for reducing high phosphorus concentrations in the lake (Fig. 7f). When phosphate (PO_4) was higher than the threshold of 0.51 ton, AWD would increase, and vice versa. Similarly, the thresholds for TP_{wd}, TN_{wd}, benthic PO_4 exchange, and wind speed were 0.066 mg L^{-1}, 2.4 mg L^{-1}, 0.028 ton d^{-1}, and 2.1 m s^{-1}, respectively. However, other factors may influence these thresholds due to their significant interaction (Fig. 7i–m, and Figs. S2, S3). For example, the higher the algae concentration in a lake, the greater the phosphorus uptake by the algae; therefore, the PO_4 mass and its SHAP values were lower (Fig. 7i). Similar interactions are commonly present in lake systems, which is the primary reason why DWDO took so many influencing factors into account.

3.4. Impact of hyperparameters on DWDO performance

The well-performing DWDO benefited from the rational architecture and hyperparameter settings. However, these hyperparameters lacked a uniform standard, and their values were based on experiences and experimentation, such as deep and reinforcement learning [43]. We selected the important hyperparameters in DWDO and conducted the scenario analysis in order to identify the sensitivity of DWDO, and to know how they alter the decision (Table 3).

The effect of multiple hyperparameters on the mean reward of the episodes is shown in Fig. 8. In scenario A3, DWDO converged quickly to a high reward when the weight of the water quality reward (k_{wq}) was the highest, and the lake had a high target of TP (Fig. 8a). Scenario A4 was similar to A3 when converging on the 50 episodes. After the k_{wq} decreased, scenario A5 took 100 episodes to converge, and the converged reward became increasingly lower. Therefore, increasing the k_{wq} was helpful to increase the reward. If the k_{wq} were constant and the water quality targets changed, the reward and convergence speed were higher in scenarios A3–A6 than in A1 and A2 (likewise, B2 is higher than B1 in Fig. 8b). We found the water quality targets strongly influenced the reward function. Assuming that the TN_{wd} or TP_{wd} was higher than the lake at any given time; the more water was transferred, the more pollutants entered the lake, and the TN or TP of the lake would be worse even though its water circulation is faster. If the TN_{wd} or TP_{wd} was close to the threshold WQ_{wd0}, changing the AWD does not significantly improve the water quality of the lake (Figs. S4, S5). Therefore, the water quality target must be lower than WQ_{wd0}. Otherwise, the larger the AWD, the worse the lake water quality. Furthermore, the WQ_{wd0} might change over time. In this study, we set $wq_{TN,target}$ as 2 mg L^{-1} and $wq_{TP,target}$ as 0.05 mg L^{-1} and were consistent with Class V and III of the Chinese Environmental Quality Standards for surface water, respectively, which were below the dynamic changing thresholds.

The DDPG calculated the cumulative reward over multiple time steps using the discount factor γ. The smaller γ is, the faster the next reward declines [44]. Additionally, the higher the reward, the faster the training (Fig. 8c). The training interval time and the soft update factor of the networks had little effect on the DWDO reward (Fig. 8d and e); however, the network architecture and learning rate of the optimizer affects the stability of the algorithm (Fig. 8f and Fig. S6). Remarkably, the DWDO failed to train in scenarios A2 and F1, meaning that the reward suddenly decreased instead of increasing during training. It was inferred that the random initialization of DDPG and the stochastic process (TN_{wd} or TP_{wd} from a uniform distribution) might have caused this problem. It is noteworthy that, unsuitable hyperparameters may affect the convergence of the DRL or cause non-convergence.

3.5. Visions of integrating DRL with water quality management

As mentioned above, deep reinforcement learning maximally optimized the dynamic water diversion decision problem through interactions with the *Environment*. As a general and expansible approach, it decouples the dynamic optimization problem into two systems that interact sequentially over time. One is an entity (agent) that makes decisions, and the other is an object (*Environment*) affected by those decisions. Any factor outside of the entity can be part of the *Environment*. This architecture enables DRL to cope with various optimization problems, even if the states of the *Environment* are partially observable in the simulated or real world. The optimal strategy learned by the agent results from the systematic evolution of perception, knowledge representation,

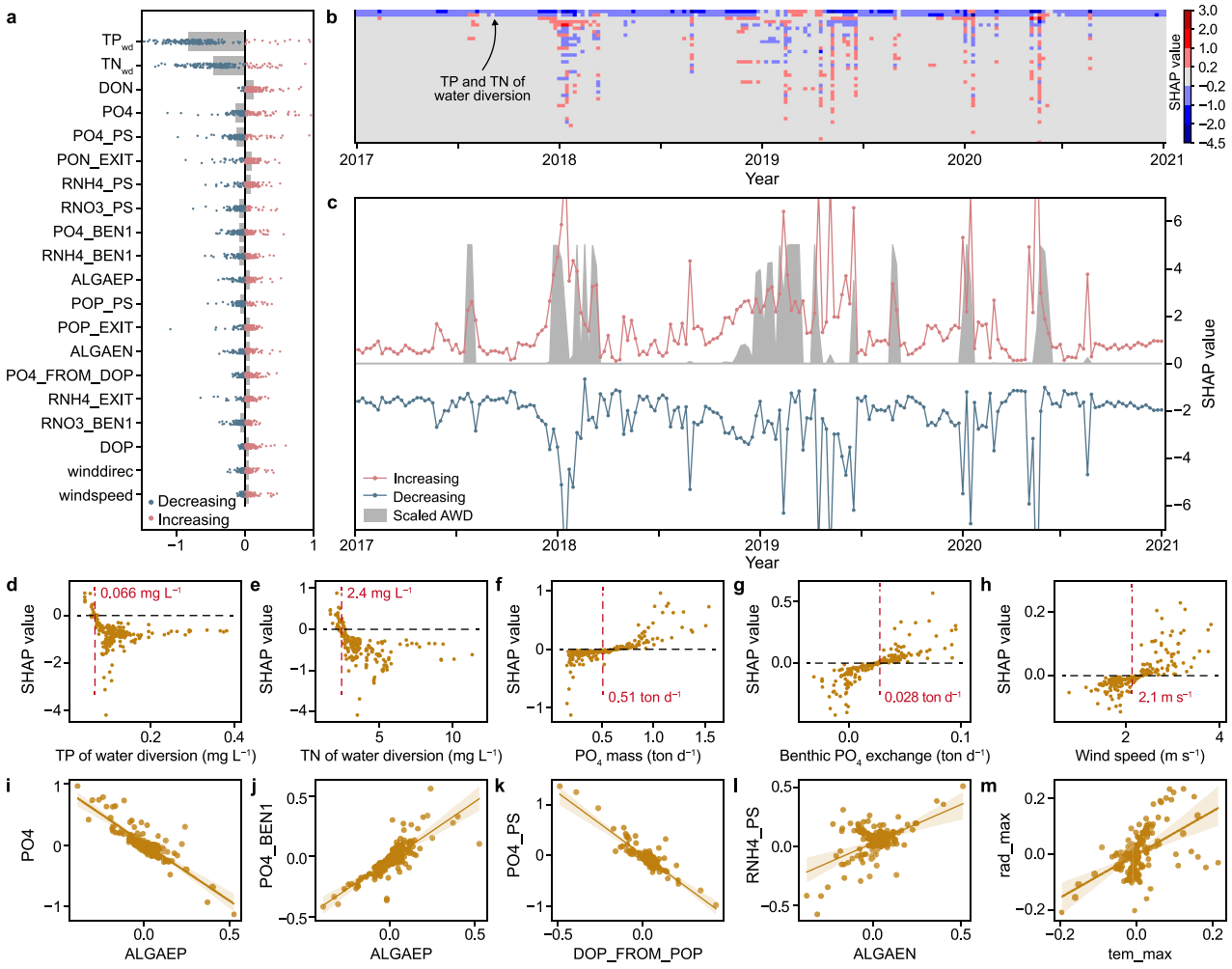

Fig. 7. The contribution (SHAP value) of input factors to the diversion decision and their interactions. **a**, The contribution distribution of factors to increase or decrease the decision. The bar represents the average contribution, and the y-axis represents the contribution of TP and TN concentration of water diversion (TP_{wd} and TN_{wd}), Dissolved organic nitrogen mass (DON), PO_4 mass (PO4), PO_4 input (PO4_PS), PO_4 outflow (PO4_EXIT), NH_3—N input (RNH4_PS), NO_3—N input (RNO3_PS), Benthic PO_4 exchange (PO4_BEN1), Benthic NH_3—N exchange (RNH4_BEN1), Phosphorus in algae (ALGAEP), Particle organic phosphorus input (POP_PS), Particle organic phosphorus outflow (POP_EXIT), Nitrogen in algae (ALGAEN), Mineralization of DOP (PO4_FROM_DOP), NH_3—N outflow (RNH4_EXIT), Benthic NO_3—N interaction (RNO3_BEN1), Dissolved organic phosphorus mass (DOP), Average hourly wind speed (windspeed), Average hourly wind direction (winddirec). **b**, The heatmap of contribution along a timeline. The x-axis represents the timeline, and the y-axis represents 67 input factors (Table S3). The redder the color in the heatmap represents a greater contribution to an increase in AMD. The bluer the color represents a greater contribution to a decrease in AMD. **c**, The positive (increasing the AWD) and negative (decreasing the AWD) contribution trends and their net value (corresponding to AWD). **d**–**h**, The relationships of SHAP values and TP of water diversion (**d**), TN of water diversion (**e**), PO_4 mass of Lake Caohai (**f**), benthic PO_4 exchange (**g**), and wind speed (**h**). The red dashed line depicts the threshold of the *x*-axis whose SHAP value was close to 0. **i**–**m**, The interaction of SHAP values between PO_4 mass (PO_4) and phosphorus mass in algae (ALGAEP) (**i**), benthic PO_4 exchange (PO4_BEN1) and algal biomass (ALGAEP) (**j**), PO_4 input (PO4_PS) and hydrolysis flux of POP (DOP_FROM_POP) (**k**), NH_3—N input (RNH4_PS) and algal biomass (ALGAEN) (**l**), and solar radiation (rad_max) and maximum air temperature (tem_max) (**m**).

memory, planning, imagination, and other abilities. In this study, in order to increase the water quality and water diversion rewards, an agent must equip with the ability of perception (to examine the water quality status of the lake), knowledge (to understand water quality response relationships), attention (to focus on key driving factors), memory (to remember the historical experience), and planning (to make the water diversion decision). Fortunately, the general goal of maximizing rewards was enough to drive actions that exhibit most of these abilities [45]. In addition, the agent was placed in a corresponding complex *Environment* where multiple intelligences would be reinforced. As a consequence, DRL becomes more capable of complex dynamic optimization problems [46].

The accuracy of DWDO was attributed to the process-based model and DDPG. The EFDC model was based on hydrodynamic and biochemical processes and could simulate eutrophic lakes. Assuming an extreme rainfall event and large amounts of water and non-point source pollutants were discharged into the lake, the model could simulate the adaptive adjustment of internal cycles through sedimentation, outflow, and denitrification [36]. The parameters determined the modeling performance and quantified the process rate of change. To identify the impact of the model parameters on EFDC and DWDO, we tested the variation of simulations of water quality and the optimal water diversion decisions under multiple sets of parameters (Fig. 9). Interestingly, the trends during simulated water quality and optimal decisions remained consistent with the baseline, even though the variation in parameters was large (Table S3). Furthermore, it was found that both the reward and objective functions converge to the same level during DWDO training, confirming the strong robustness of the algorithm under model parameter uncertainty.

As an integrated approach, DWDO was similar to the traditional "simulation-optimization" methods that coupled machine learning

Table 3
Hyperparameter settings in DWDO under different scenarios.

Scenario group	ID	Hyperparameters
A	A1	$k_{wq} = 95$; $k_d = 5$; $wq_{TN,target} = 2$; $wq_{TP,target} = 0.05$
	A2	$k_{wq} = 90$; $k_d = 10$; $wq_{TN,target} = 2$; $wq_{TP,target} = 0.05$
	A3	$k_{wq} = 99$; $k_d = 1$; $wq_{TN,target} = 1.5$; $wq_{TP,target} = 0.1$
	A4	$k_{wq} = 95$; $k_d = 5$; $wq_{TN,target} = 1.5$; $wq_{TP,target} = 0.1$
	A5	$k_{wq} = 90$; $k_d = 10$; $wq_{TN,target} = 1.5$; $wq_{TP,target} = 0.1$
B	B1	Action duration = 15; $wq_{TN,target} = 2$; $wq_{TP,target} = 0.05$
	B2	Action duration = 15; $wq_{TN,target} = 1.5$; $wq_{TP,target} = 0.1$
C	C1	$\gamma = 0.99$
	C2	$\gamma = 0.95$
	C3	$\gamma = 0.90$; $wq_{TN,target} = 1.5$; $wq_{TP,target} = 0.1$
D	D1	Training interval = 5
	D2	Training interval = 10
E	E1	Target model update rate = 0.005
	E2	Target model update rate = 0.01
F	F1	Learning rate = 0.005
	F2	Learning rate = 0.01

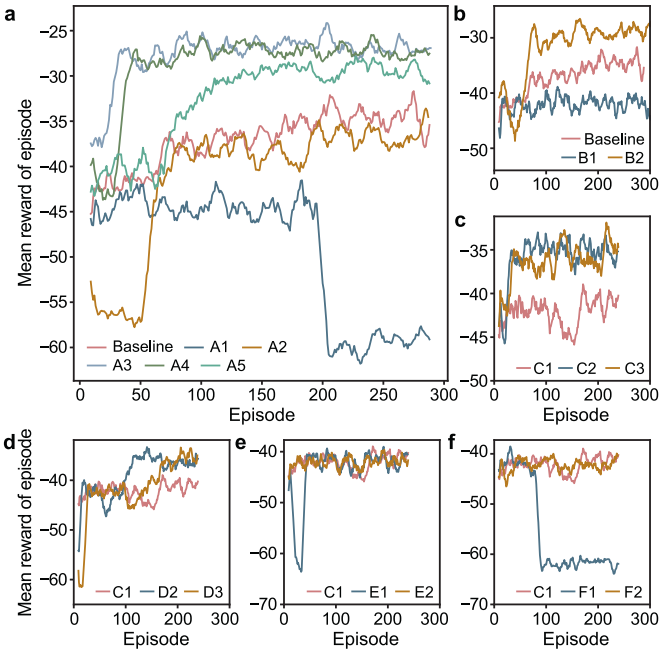

Fig. 8. The impact of different DWDO hyperparameters on the mean reward of each episode. **a**, The weights of the reward and water quality targets; **b**, Action duration; **c**, Discount factor; **d**, Training interval; **e**, Target networks update rate; **f**, Learning rate.

algorithms to a mechanistic model [47–49]. In general, the "simulation-optimization" method consumes a high computational cost due to the numerical solution of hydrodynamic, water quality, and algal equations of thousands of grids over millions of time steps. Deep learning training is another time-consuming task in DRL; however, it balances underfitting and overfitting [50]. In this study, DWDO took only three days to complete the training on a personal computer with an 8-core CPU (Intel i7-10700) and 32 GB RAM. The fast training was attributed to the following facts: (a) DRL stored and reused recent historical data, so the usage of the data was higher than the "simulation-optimization" methods; (b) DRL simplified the optimization of maximizing cumulative rewards of multiple time steps into the calculation of reward through the discount factor, and had it update the neural network weights at each step. Nevertheless, the "simulation-optimization" methods typically optimize the decision at the end of the entire period [49,51]. In addition, DRL handled the trade-off between

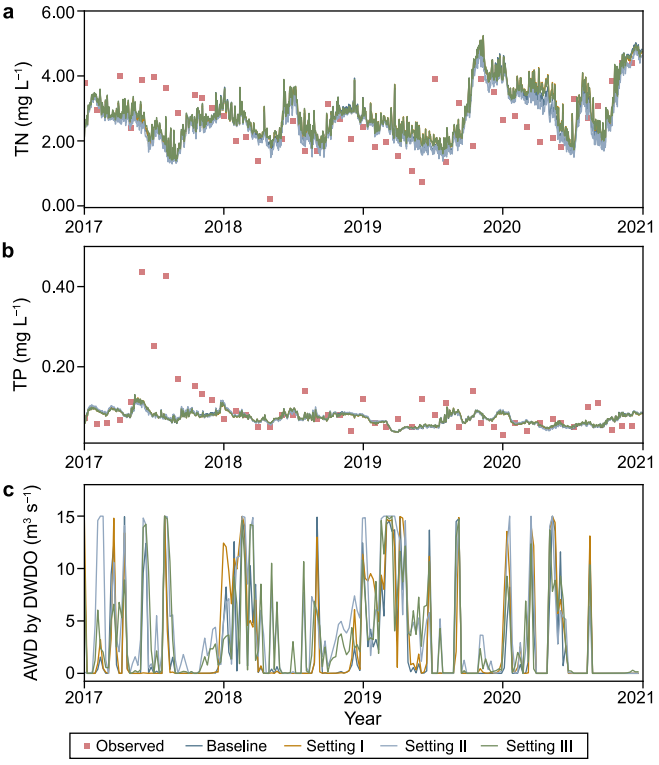

Fig. 9. The change of simulated TN (**a**), simulated TP (**b**), and optimal amount (**c**) of water diversion by DWDO, under the four settings of uncertain EFDC parameters that were recorded in Table S3.

exploitation and exploration better than common optimization algorithms and was robust in searching for the global optimum [52]. For example, in this study, DRL did not directly optimize the AWD, instead optimized the massive weights of DL, which stood for the AWD generating strategy. This strategy can be understood as an approach to decision-making rather than the specific values of the decision variables used by traditional algorithms.

This provided new ideas for complex optimization problems and can be used to solve long-lasting optimization problems in a static system. For instance, the problem of the annual compliance goal of water quality will be decomposed into multiple subgoals at an interval of one month. We then used DRL to solve and obtain the optimal operating decisions for each month. In addition, DRL is better at solving the following two types of problems: (a) sparse reward (delayed reward feedback) and diverse types of state and action spaces [53,54]; and (b) decision-making of multiple agents in game theory [30,55,56]. Here an agent will regard the action of other agents as the states of the *Environment*. These advantages have greatly progressed in robotics, industrial automation, and driverless cars. Therefore, DRL will make a significant difference in water quality management.

In summary, the systematic evolution of intelligent abilities, including perception, knowledge representation, memory, and planning, dominate DWDO performance within a complex *Environment*. DWDO outweighed the traditional methods with a faster training speed and has great potential in water quality management.

4. Conclusions

To maximize water quality improvement of water diversion and reduce operation costs, we proposed a dynamic DWDO. Using a

coupled complex process-based model with deep reinforcement learning, we verified its performance on a eutrophic lake, explained optimal decision-making strategies, and tested the hyperparameters' diverse sets. The main conclusions are:

(a) The Niulan River—Dianchi Water Diversion project aimed to improve the water quality of the eutrophic lake; however, its effectiveness was vulnerable to the changing water quality and the lake, as well as meteorological factors.
(b) The DWDO training was conducted using hundreds of episodes via a multi-stage strategy trial, leading to decreased TN and TP of the lake by 7% and 6%, respectively, and a dramatic reduction of AWD.
(c) The contribution of states to the diversion decision varied with the specific value of inputs. The adaptive adjustment of the diversion decision was dominated by the water quality of water diversion and the interactions between the input factors that co-drove the change of AWD.
(d) Hyperparameters in the *Environment* and DDPG had a significant impact on the reward convergence of DWDO. It was suitable to meet diverse preferences on the weights of water quality and diversion, water quality targets, and action duration. Meanwhile, DWDO was robust under various model parameter uncertainties.

CRediT authorship contribution statement

Qingsong Jiang: Conceptualization, Resources, Investigation, Methodology, Writing - Original Draft. **Jincheng Li:** Resources, Data Curation, Investigation, Formal Analysis. **Yanxin Sun:** Data Curation, Formal Analysis. **Jilin Huang:** Data Curation, Formal Analysis. **Rui Zou:** Conceptualization, Methodology. **Wenjing Ma:** Resources, Methodology, Formal Analysis. **Huaicheng Guo:** Supervision, Conceptualization. **Zhiyun Wang:** Investigation, Data Curation. **Yong Liu:** Supervision, Conceptualization, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was financially supported by the National Social Science Foundation of China (21AZD060), China; the National Natural Science Foundation of China (51721006), China and the High-Performance Computing Platform of Peking University, China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2023.100298.

References

[1] S. Wang, J. Li, B. Zhang, E. Spyrakos, A.N. Tyler, Q. Shen, F. Zhang, T. Kuster, M.K. Lehmann, Y. Wu, Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index, Remote Sens. Environ. 217 (2018) 444–460.
[2] J. Chen, T. Zhang, P. Du, Assessment of water pollution control strategies: a case study for the Dianchi Lake, J. Environ. Sci. 14 (1) (2002) 76–78.
[3] L. Carvalho, C. McDonald, C. de Hoyos, U. Mischke, G. Phillips, G. Borics, S. Poikane, B. Skjelbred, A.L. Solheim, J. Van Wichelen, Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control, J. Appl. Ecol. 50 (2) (2013) 315–323.
[4] C.J. Stevens, Nitrogen in the environment, Science 363 (6427) (2019) 578–580.
[5] J.C. Sheng, W.Z. Tang, Spatiotemporal variation patterns of water pollution drivers: the case of China's south-north water transfer project, Sci. Total Environ. 761 (2021) 143190.
[6] L.J. Zhang, J.H. Yang, Y. Zhang, J.Z. Shi, H.X. Yu, X.W. Zhang, eDNA biomonitoring revealed the ecological effects of water diversion projects between Yangtze River and Tai Lake, Water Res. 210 (2022) 117994.
[7] C.Y. Tang, C. He, Y.P. Li, K. Acharya, Diverse responses of hydrodynamics, nutrients and algal biomass to water diversion in a eutrophic shallow lake, J. Hydrol. 593 (2021) 125933.
[8] R.Y. Zhang, B.S. Wu, Environmental impacts of high water turbidity of the Niulan River to Dianchi lake water diversion project, J. Environ. Eng. 146 (1) (2020) 05019006.
[9] Y. Li, C. Tang, C. Wang, D.O. Anim, Z. Yu, K. Acharya, Improved Yangtze River diversions: are they helping to solve algal bloom problems in Lake Taihu, China? Ecol. Eng. 51 (2013) 104–116.
[10] X.L. Zhang, R. Zou, Y.L. Wang, Y. Liu, L. Zhao, X. Zhu, H.C. Guo, Is water age a reliable indicator for evaluating water quality effectiveness of water diversion projects in eutrophic lakes? J. Hydrol. 542 (2016) 281–291.
[11] R.I. Woolway, B.M. Kraemer, J.D. Lenters, C.J. Merchant, C.M. O'Reilly, S. Sharma, Global lake responses to climate change, Nat. Rev. Earth Environ. 1 (8) (2020) 388–403.
[12] B.Q. Qin, G. Gao, G.W. Zhu, Y.L. Zhang, Y.Z. Song, X.M. Tang, H. Xu, J.M. Deng, Lake eutrophication and its ecosystem response, Chin. Sci. Bull. 58 (9) (2013) 961–970.
[13] E. Jeppesen, B. Kronvang, J.E. Olesen, J. Audet, M. Sondergaard, C.C. Hoffmann, H.E. Andersen, T.L. Lauridsen, L. Liboriussen, S.E. Larsen, M. Beklioglu, M. Meerhoff, A. Ozen, K. Ozkan, Climate change effects on nitrogen loading from cultivated catchments in Europe: implications for nitrogen retention, ecological state of lakes and adaptation, Hydrobiologia 663 (1) (2011) 1–21.
[14] R.D. Gulati, E. van Donk, Lakes in The Netherlands, their origin, eutrophication and restoration: state-of-the-art review, Hydrobiologia 478 (1–3) (2002) 73–106.
[15] M. Farina, K. Deb, P. Amato, Dynamic multiobjective optimization problems: test cases, approximations, and applications, IEEE Trans. Evol. Comput. 8 (5) (2004) 425–442.
[16] A.R. Jordehi, Particle swarm optimisation for dynamic optimisation problems: a review, Neural Comput. Appl. 25 (7–8) (2014) 1507–1516.
[17] E. Kim, H.G. Kim, S. Baek, M. Cho, Effective structural optimization based on equivalent static loads combined with system reduction method, Struct. Multidiscip. Optim. 50 (5) (2014) 775–786.
[18] R.S. Sutton, TD Models: modeling the world at a mixture of time scales, in: A. Prieditis, S. Russell (Eds.), Machine Learning Proceedings, Morgan Kaufmann, San Francisco (CA, 1995, pp. 531–539.
[19] M. Rothmann, M. Porrmann, A survey of domain-specific architectures for reinforcement learning, IEEE Access 10 (2022) 13753–13767.
[20] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robot. 37 (3) (2020) 362–386.
[21] W.S. Zhao, J.P. Queralta, T. Westerlund, Ieee, sim-to-real transfer in deep reinforcement learning for robotics: a survey, in: IEEE Symposium Series on Computational Intelligence (IEEE SSCI), Ieee: Electr Network, 2020, pp. 737–744.
[22] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, D. Silver, Mastering Atari, Go, chess and shogi by planning with a learned model, Nature 588 (7839) (2020) 604–609.
[23] T.T. Nguyen, N.D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications, IEEE Trans. Cybern. 50 (9) (2020) 3826–3839.
[24] J.H. Lee, J.W. Labadie, Stochastic optimization of multireservoir systems via reinforcement learning, Water Resour. Res. 43 (11) (2007) W11408.
[25] A. Castelletti, S. Galelli, M. Restelli, R. Soncini-Sessa, Tree-based reinforcement learning for optimal water reservoir operation, Water Resour. Res. 46 (2010).
[26] A. Castelletti, H. Yajima, M. Giuliani, R. Soncini-Sessa, E. Weber, Planning the optimal operation of a multioutlet water reservoir with water quality and quantity targets, J. Water Resour. Plann. Manag. 140 (4) (2014) 496–510.
[27] K. Madani, M. Hooshyar, A game theory-reinforcement learning (GT-RL) method to develop optimal operation policies for multi-operator reservoir systems, J. Hydrol. 519 (2014) 732–742.
[28] A. Castelletti, F. Pianosi, M. Restelli, A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run, Water Resour. Res. 49 (6) (2013) 3476–3486.
[29] L. Lu, H. Zheng, J. Jie, M. Zhang, R. Dai, Reinforcement learning-based particle swarm optimization for sewage treatment control, Complex Intell. Syst. 7 (5) (2021) 2199–2210.
[30] K.H. Chen, H.C. Wang, B. Valverde-Perez, S.Y. Zhai, L. Vezzaro, A.J. Wang, Optimal control towards sustainable wastewater treatment plants based on multi-agent reinforcement learning, Chemosphere 279 (2021) 130498.
[31] P. Zhou, X. Wang, T.Y. Chai, Multiobjective operation optimization of wastewater treatment process based on reinforcement self-learning and knowledge guidance, IEEE Trans. Cybern. (2022) 1–14.
[32] B.D. Bowes, C. Wang, M.B. Ercan, T.B. Culver, P.A. Beling, J.L. Goodall, Reinforcement learning-based real-time control of coastal urban stormwater systems to mitigate flooding and improve water quality, Environmental

Science-Water Research & Technology 8 (10) (2022) 2065–2086.
[33] S.M. Saliba, B.D. Bowes, S. Adams, P.A. Beling, J.L. Goodall, Deep reinforcement learning with uncertain data for real-time stormwater system control and flood mitigation, Water 12 (11) (2020) 3222.
[34] Z. Wang, R. Zou, X. Zhu, B. He, G. Yuan, L. Zhao, Y. Liu, Predicting lake water quality responses to load reduction: a three-dimensional modeling approach for total maximum daily load, Int. J. Environ. Sci. Technol. 11 (2) (2014) 423–436.
[35] H.L. Hu, Reward and aversion, in: S.E. Hyman (Ed.), Annual Review of Neuroscience, vol. 39, Annual Reviews, Palo Alto, 2016, pp. 297–324.
[36] N.N. Ji, R. Zou, Q.S. Jiang, Z.Y. Liang, M.C. Hu, Y. Liu, Y.H. Yu, Z.Y. Wang, H.L. Wang, Internal positive feedback promotes water quality improvement for a recovering hyper-eutrophic lake: a three-dimensional nutrient flux tracking model, Sci. Total Environ. 772 (2021) 145505.
[37] C.C. Carey, E. Rydin, Lake trophic status can be determined by the depth distribution of sediment phosphorus, Limnol. Oceanogr 56 (6) (2011) 2051–2063.
[38] C.R. Qiu, Y. Hu, Y. Chen, B. Zeng, Deep deterministic policy gradient (DDPG)-Based energy harvesting wireless communications, IEEE Internet Things J. 6 (5) (2019) 8577–8588.
[39] K.G. Kapanova, I. Dimov, J.M. Sellier, A genetic approach to automatic neural network architecture optimization, Neural Comput. Appl. 29 (5) (2018) 1481–1492.
[40] Z.J. Zhang, Improved Adam optimizer for deep neural networks, in: IEEE International Symposium on Quality of Service, IWQOS, 2018, pp. 1–2.
[41] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: 31st Annual Conference on Neural Information Processing Systems (NIPS), vol. 30, Neural Information Processing Systems (Nips), Long Beach, CA, 2017.
[42] K.R. Jin, Z.G. Ji, Case study: modeling of sediment transport and wind-wave impact in Lake Okeechobee, J. Hydraul. Eng. 130 (11) (2004) 1055–1067.
[43] R. Liessner, J. Schmitt, A. Dietermann, B. Baker, Hyperparameter optimization for deep reinforcement learning in vehicle energy management, in: 11th International Conference on Agents and Artificial Intelligence (ICAART), Scitepress, Prague, CZECH REPUBLIC, 2019, pp. 134–144.
[44] R. Amit, R. Meir, K. Ciosek, Discount factor as a regularizer in reinforcement learning, in: International Conference on Machine Learning (ICML), Jmlr-Journal Machine Learning Research, vol. 119, Electr Network, 2020.
[45] D. Silver, S. Singh, D. Precup, R.S. Sutton, Reward is enough, Artif. Intell. 299 (2021) 13.
[46] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science 362 (6419) (2018) 1140–1144.
[47] F.F. Dong, Z.Z. Zhang, C. Liu, R. Zou, Y. Liu, H.C. Guo, Towards efficient Low Impact Development: a multi-scale simulation-optimization approach for nutrient removal at the urban watershed, J. Clean. Prod. 269 (2020) 122295.
[48] M. Asadzadeh, S. Razavi, B.A. Tolson, D. Fay, Pre-emption strategies for efficient multi-objective optimization: application to the development of Lake Superior regulation plan, Environ. Model. Software 54 (2014) 128–141.
[49] C. Dai, Q. Tan, W.T. Lu, Y. Liu, H.C. Guo, Identification of optimal water transfer schemes for restoration of a eutrophic lake: an integrated simulation-optimization method, Ecol. Eng. 95 (2016) 409–421.
[50] L. Ali, A. Rahman, A. Khan, M.Y. Zhou, A. Javeed, J.A. Khan, An automated diagnostic system for heart disease prediction based on chi(2) statistical model and optimally configured deep neural network, IEEE Access 7 (2019) 34938–34945.
[51] A. Al-Maktoumi, M.M. Rajabi, S. Zekri, C. Triki, A probabilistic multiperiod simulation-optimization approach for dynamic coastal aquifer management, Water Resour. Manag. 35 (2021) 3447–3462.
[52] S. Ishii, W. Yoshida, J. Yoshimoto, Control of exploitation-exploration meta-parameter in reinforcement learning, Neural Network. 15 (4–6) (2002) 665–687.
[53] J. Andreas, D. Klein, S. Levine, Modular multitask reinforcement learning with policy sketches, in: 34th International Conference on Machine Learning, vol. 70, Jmlr-Journal Machine Learning Research, Sydney, AUSTRALIA, 2017.
[54] C. Wang, J. Wang, J.J. Wang, X.D. Zhang, Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards, IEEE Internet Things J. 7 (7) (2020) 6180–6190.
[55] W.C. Tian, Z.L. Liao, G.Z. Zhi, Z.Y. Zhang, X. Wang, Combined sewer overflow and flooding mitigation through a reliable real-time control based on multi-reinforcement learning and model predictive control, Water Resour. Res. 58 (7) (2022) e2021WR030703.
[56] F.W. Hung, Y.C.E. Yang, Assessing adaptive irrigation impacts on water scarcity in nonstationary environments-A multi-agent reinforcement learning approach, Water Resour. Res. 57 (9) (2021) e2020WR029262.