# CSO Reduction through a Safe Real-Time Control based on Multi-Reinforcement Learning, system optimization, and Model Predictive Control

Zhenliang Liao[1,2,3,4,5], Wenchong Tian[2,3†], Guozheng Zhi[6], Xuan Wang[2,3].

[1] College of Civil Engineering and Architecture, Xinjiang University, 830046, Urumqi, China.

[2] State Key Laboratory of Pollution Control and Resource reuse, College of Environmental Science and Engineering, Tongji University, 200092 Shanghai, China.

[3] Key Laboratory of Yangtze River Water Environment, Ministry of Education, Tongji University, 200092 Shanghai, China.

[4] Shanghai Institute of Pollution Control and Ecological Security, Shanghai 200092, P.R. China.

[5] UNEP-Tongji Institute of Environment for Sustainable Development, College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China.

†Corresponding author: Wenchong Tian

[6] Shanghai Urban Water Resources Development and Utilization National Engineering Center Co. L td., Shanghai 200082, China

Email address: wenchong@tongji.edu.cn

Tel: +86 18817870656

Abstract

Real-time control (RTC) helps the combined sewer system to adapt its response to individual rainfall and enhance the performance of combined sewer overflow (CSO) reduction. Recently, an RTC approach based on reinforcement learning (RL) is developed for flooding control in a stormwater system. However, the safety and the performance of this AI algorithm still need further improvement. In this paper, a new RTC method based on multiple RLs, system optimization, and model predictive control (MPC) is developed for the improvement of both safety and CSO reduction. First, five RL agents are trained by five individual RL algorithms. Then, an optimization model is used to optimize the advantage function of all the agents for control effect improvement. After that, an MPC-based security system is established to check the safety of control strategy before the implementation. Finally, our new RTC model, called voting system, is developed through the combination of these five agents and the security system. This method is evaluated in the combined sewer system model of a city in eastern China. According to the results: (i) All the five trained RLs are able to show promise in overflow reduction. (ii) The AFI improves the CSO reduction of all the agents with the maximum improvement rate of 44.5%. (iii) The security system selects a safe control strategy through a small scale MPC, thus it provides a guarantee of safety. Still, our method faces the challenges of computing time, local optimization, and the limitation of system capacity.

**Key words:** safety, reinforcement learning, combined sewer system, real-time control, model predictive control, combined sewer overflow.

## 1. Introduction

The combined sewer system is widely used in many cities around the world. However, the combined sewer overflow (CSO) cannot be easily avoided during application (Mailhot et al., 2015; Suarez and Puertas, 2005; Wan and Lemmon, 2007; Xu and Liao, 2013; Xie et al., 2017; Gu et al., 2017). A primary solution to this problem is to enhance the infrastructure for optimized system-scale performance. But the cost and viability of this solution is highly variable in different cases (Abhiram et al., 2020).

Another solution is real-time control (RTC), which uses sensor data to infer the real-time state of a combined sewer system and responds via control of distributed control assets, such as valves, gates, and pumps (Rauch and Harremoes, 1999. Kerkez et al., 2016; Lund et al., 2018; Lund et al., 2020). By achieving system-level coordination between many distributed control elements, only a small set of infrastructures are needed to optimize system operation for lower overflow and flooding (Schütze et al., 2002; Kerkez et al., 2016).

Recently, some researchers provided a new type of RTC methods based on reinforcement learning (RL) and proved that it is capable of control a stormwater system in real-time for flow control and flooding reduction (Ochoa et al., 2019; Abhiram et al., 2020). However, challenged by the risk of handing over the control process to a computer, it presently shows that an independent security system is strongly demanded to guarantee the safety of the RL method in real-world operation. Meanwhile, it is still necessary to further improve the control effect of the RL method in the combined sewer system.

Considering the improvement of both the safety and the control effect of the existing RL system, a new RTC method based on multi-RL, system optimization, and model predictive control (MPC) is developed in this study. First, five RL algorithms are used to train five individual agents. Then an optimization model of combined sewer system is employed to optimize the advantage function of these RL agents, thus further improves their control effect. After that, an independent security system, which is based on a small scale MPC, is established to check the safety of control strategy before implementing it. Finally, our new RTC method, called voting system, is established through the combination of these five RL agents and the security system. Accordingly, the contributions of this paper include: 1. Using multiple RL methods, including both value-based and policy-based, to illustrate the effectiveness of different RL models on CSO reduction; 2. Improving the control effect of RL models through an optimization model based on the combined sewer system; 3. Designing an independent security system based on MPC to provide a guarantee of safety.

The remainder of this paper is organized as follows: In Section 2, we briefly introduce some related works, including the RTC in the urban drainage systems, the RL, and the concept of safe RL. In Section 3, we describe the details of our method. The case study is introduced in Section 4. And the use of our method in the case study

is given in Section 5. For comparison, the use of other RL methods is also provided. Considering the risk to property and public safety, the evaluation of these methods is established across a series of simulations, which span various rainfall events with a mathematical model (SWMM). In Section 6, the effect of CSO reduction, the safety of the voting system, and some remaining challenges are discussed. Finally, our conclusions are shown in Section 7.

## 2. Preliminaries and related works

### 2.1. RTC of urban drainage system and model predictive control

A combined sewer system is controlled in real time if process variables are monitored and used to operate actuators of the system (Schütze et al., 2002). In a control loop of an RTC system, the sensors monitor process variables and send it to controllers. The controllers operate actuators according to control strategy. Then, the actuators influence the process to optimize system operation (Schütze, et al., 2002).

A control strategy or “control procedure” is defined as the time sequence of set-points given by controller (Schütze, et al., 2002). To generate the control strategy, one of the efficient ways is the model predictive control (MPC, Fig.1), which recursively repeats the optimization of the control strategy based on a rainfall prediction within a finite time horizon and move forward according to the receding horizon principle (Fu et al., 2008; Joseph et al., 2015; Lund et al., 2018). Although MPC faces the problems of high computation load and uncertainty prediction, it is still widely used in many cases (Sebastian and Stefan, 2019; Lund et al., 2020; Congcong et al., 2020).

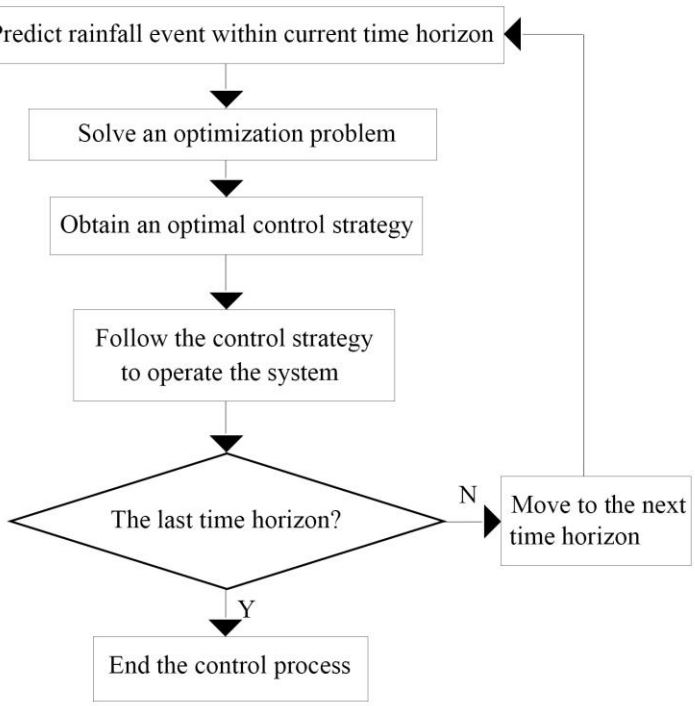



Fig.1. The schematic MPC

## 2.2. Multi-reinforcement learning

### 2.2.1. Brief review of reinforcement learning

Reinforcement learning (RL) is a kind of model used for control and planning (Sutton and Barto, 2018). The goal of RL is to learn an optimal control strategy from experimental trials and relatively simple feedback. For now, the RL has emerged as a state-of-the-art methodology for many autonomous control systems, such as autonomous driving (Pan et al., 2017), stock trading (Tan et al., 2011), AI gaming (Wu et al., 2018; Shao et al., 2018; Silver et al., 2017), reservoir scheduling systems (Madani and Hooshyar, 2014, Castelletti et al., 2013), flow control (Ochoa et al., 2019), in-line storage control (Labadie, 2014) and watershed flooding control (Abhiram et al.,2020).

Usually, an RL model includes an agent with a behavior function (policy) and an environment that is controlled by the agent. The environment provides state ($s_t$) of current time point $t$, and the agent choose to take an action ($a_t$) of next time step according to the given state and its policy. Once an action is taken, the environment delivers a reward ($r_t$) as feedback. With the help from the state, action, reward, the agent is able to master the control process and adapt to the environment actively to maximize expected future rewards (called value function, or Q value).

### 2.2.2. Multi-RL methods

Many RL models have been developed in recent years, including deep Q learning and dueling deep Q learning (Minh et al., 2015), proximal policy optimization (Schulman et al., 2017), and advance Actor-Critic (Sutton et al., 1999; Minh et al., 2016). These methods are classified as policy-based and value-based.

(1) Deep Q learning and dueling deep Q learning

Deep Q neural network learning (DQN) and dueling deep Q neural network learning (DDQN) are two methods belonging to value-based family. DQN is based on the theory of Q learning, but take advantage from deep neural network to maximum the Q value (Eq. (1), where $a_t$, $s_t$, $r_t$ are action, state, and reward according to above, $\theta$ is the parameters of neural network, $\gamma$ is a hyperparameter called the discount factor), to find the parameters enlarging the total expected reward during control process (Minh et al., 2015; Sutton and Barto, 2018).

$$q(a_t, s_t, \theta) = \mathbb{E}_t\left[\sum_{k=0}^{\infty}[\gamma^k r_{t+k+1}|a_t, s_t, \theta]\right] \quad (1)$$

DDQN uses two deep neural networks to approximate the optimal Q value (Van 2010), one of them is to determine the optimal policy and the other is to determine the Q value (Van et al., 2016; Wang et al., 2016). In this way, DDQN avoids overoptimistic value estimates problem of DQN.

(2) Proximal policy optimization

Proximal policy optimization models, including PPO1 and PPO2, are the policy-based reinforcement learning models (Schulman et al., 2015; Schulman et al., 2017). PPO1 tries to find the best policy function (input state, and output action) by computing an estimator of the policy gradient (Eq. (2), where $\pi_\theta(a_t|s_t)$ is policy with parameter $\theta$, $q$ is the Q value given by policy) and plugging it into a stochastic gradient ascent algorithm (Schulman et al., 2017).

$$g = \mathbb{E}_t[\nabla_\theta log\pi_\theta(a_t|s_t)q] \quad (2)$$

However, randomly changing happens during policy update of PPO1. To avoid this, PPO2 imports a penalty on Kullback-Leibler (KL) divergence to the clipped surrogate objective, thus it has a more stable policy updating (Schulman et al., 2017).

(3) Advance Actor-Critic

Advance Actor-Critic (A2C) combines the ingredient of both policy-based and value-based to simultaneously upgrade both policy function and maximum Q value

(Minh et al., 2016). It uses two deep neural networks, actor-learner and critic-learner to represent policy function and maximum Q value. The Actor is a reference to the learned policy function, and Critic refers to the learned Q value function. The training process of A2C aims to upgrade both Actor and Critic for a better control procedure.

2.2.3. Advantage function

According to above, the Q value (Eq. (1)) is mainly used as an evaluation of system control. In many researches, it is replaced by the advantage function $A_t$ (Eq. (3)).

$$A_t(s_t, a_t, \theta) = q(s_t, a_t, \theta) - \max_{a_t} q(s_t, a_t, \theta) \quad (3)$$

The advantage function is used to measure the difference between Q value and the estimation of maximum Q value (or value function), which represents 'how much we earn closer to the top' (Minh et al., 2016). Therefore, it provides information about the best control process. However, this information is estimated by sampled Q value, rather than an actual maximum Q value. It may be strongly influenced by the randomness of the sampling process. Thus, the advantage function can be improved when a more reliable estimation is given.

2.3.Safety and safe reinforcement learning

The safety in the RL field means to ensure reasonable system performance and respect safety constraints (Garcia and Fernandez, 2015). This definition does not necessarily refer to physical issues, as the detailed safety requirement is problem-dependent. Usually, a stable water level of the structures and the safe operation of the facilities are two reasonable safety requirements of the combined sewer system.

In the RL literature, achieving safety usually means minimizing the variance of the total expected reward (Moldovan & Abbeel, 2012), reducing the temporal differences (Gehring & Precup, 2013), and avoiding the error state (Geibel and Wysotzki, 2005). Two main types of methods have been developed to achieve the above requirements: the optimization criterion-based method, and the exploration process-based method. The first one modifies the total expected reward by taking the safety as one aspect of the reward during the training process (Garcia and Fernandez, 2015; Castro, et al., 2012; Geibel and Wysotzki, 2005). The second one uses prior knowledge to force the agent to select safe actions with higher probability in the training process (Garcia and Fernandez, 2015; Yong et al., 2012; Pablo et al., 2013).

Although these two types of methods achieved significant improvement in the RL safety, they only focus on the training process to help the agent learn to behave safely. It means that both the security and control systems are coupled in one black-box model. However, handing over the control process and safety check to a single black-box model is not a wise choice in the real-world application, especially in civil engineering (Abhiram et al., 2020). Accordingly, a security system that is independent of the control system is necessary for real-world operation.

## 3. Methodology

A new RTC method is established through the combination of multi-RL, MPC, and system optimization in this section for the improvement of safety and efficiency. First, five RL agents are trained individually through five RL models with an environment (a combined sewer system model). Then, a new advantage function based on an optimization model is given to optimize all the RL agents for a better control effect. Meanwhile, an independent MPC-based security system is designed for safety check. Finally, the new RTC method is established through the combination of these trained agents and the independent security system. This given RTC method is used for CSO reduction in the combined sewer system. The route map is given as Fig.2.

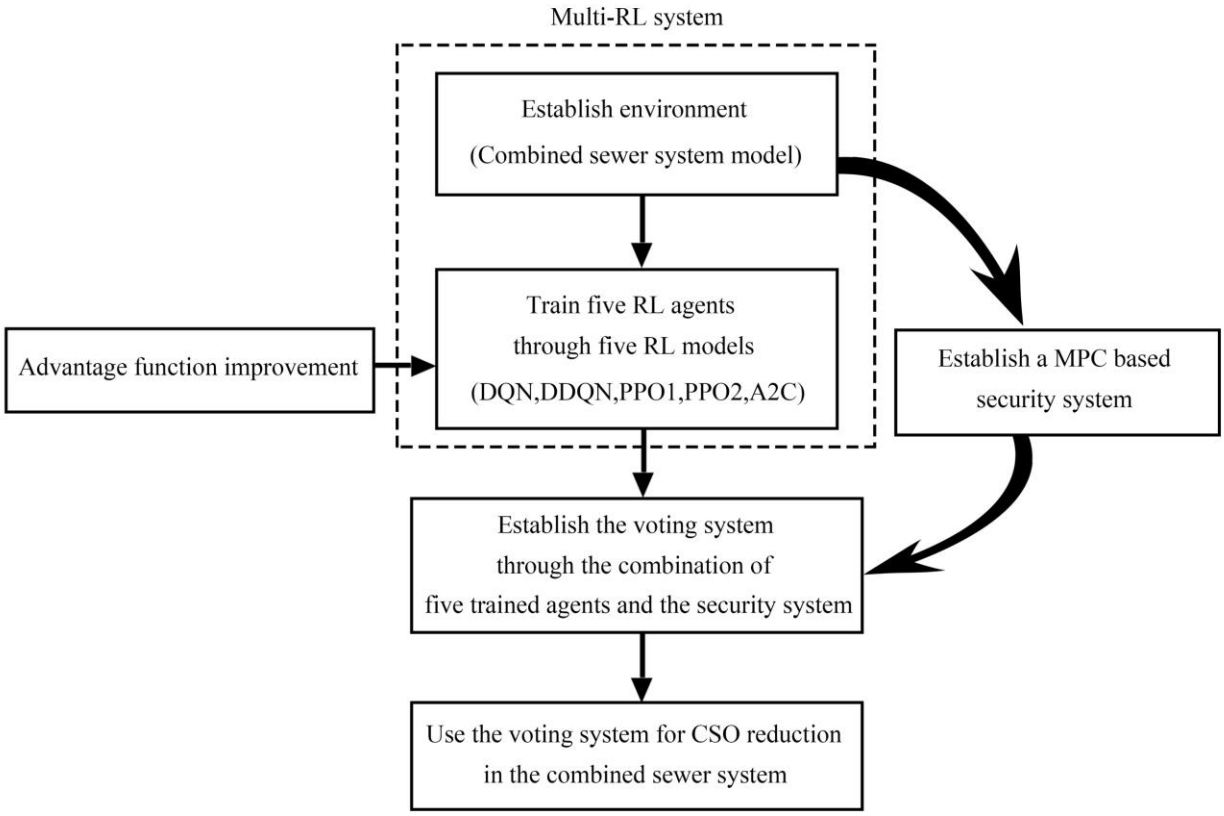



Fig.2 Route map

### 3.1. Multi-RLs based RTC for CSO reduction

We use multi-RL models to develop RTC systems for CSO reduction in the combined sewer system. Similar to previous research (Abhiram et al., 2020), the RL control systems can be described by an agent and environment. The environment represents a combined sewer system and the agent represents the entity controlling the system (Abhiram et al., 2020). During a rainfall event, an RL agent, or controller observes the state of the environment and coordinates the actions of the control assets in real-time to achieve benefits, or reward, which is the reduction of CSO in this case. The state $s_t$ can be set as an array variable which represents useful information of sewer system at time point $t$. The action $a_t$ is the control strategy of next time step given by agent with respect to the state. It can be set as a variable which indicates the operation of control assets (such as pumps, valves, gates) from time point $t$ to time point $t+1$. The reward $r_t$ is a variable which represents the training target. For instance, the agent could receive positive reward for preventing CSO or a negative reward for causing CSO.

After establishing the agent-environment system, five RL models (including DQN, DDQN, PPO1, PPO2, and A2C) are used to train five agents separately. Although the training algorithms of each RL model are different, their basic steps are similar and can be described as follow. First, we need to collect the data of states, rewards, and actions by running the agent-environment system. It means using the agent (maybe un-trained) to control the combined sewer system model under some given rainfall events, and then collecting the state-reward-action during the process. This step is called sampling. After several rounds of sampling, the collected data is used to upgrade the agent via one of the above RL models. This step is called upgrading. The collected data contains information about the environment and our expectation of control, thus the upgrading is capable of improving the agent. The system keeps running the loop of sampling-upgrading with different rainfall events until the agent achieves a good enough control effect. Repeat these steps for five RL models, then we have five trained RL agents. This training process is shown in Fig.3.

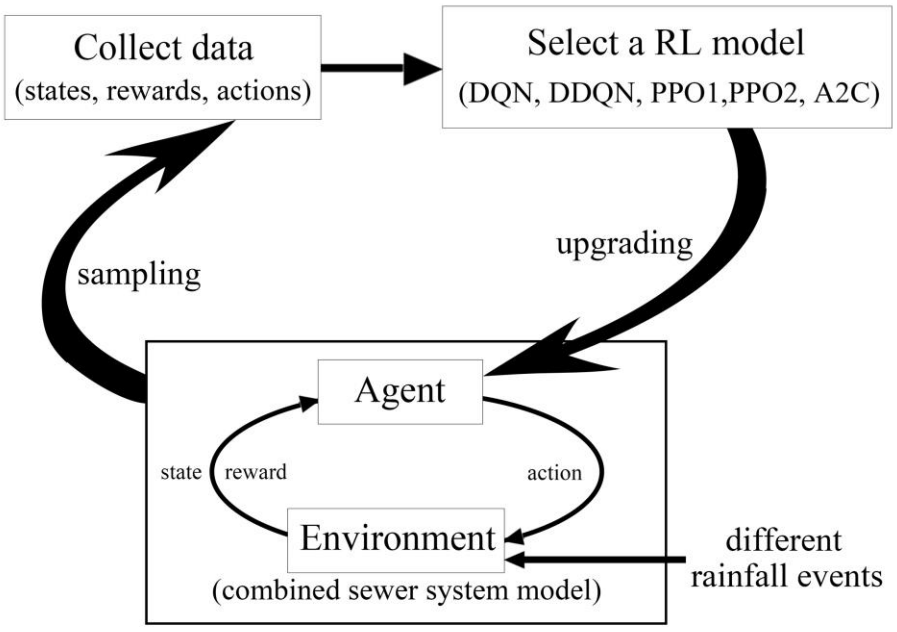



Fig.3 Training process of multi-RL models

It is impossible to hand over the control of a real-world combined sewer system directly to an untrained agent. Therefore, a simulation-based scenario is strongly needed for training (Abhiram et al., 2020). We use a well-calibrated combined sewer system model (such as an SWMM model) and some rainfall events data as a prepared virtual environment.

### 3.2. Advantage function improvement

According to section 2.2.3, a better estimation of the maximum Q value can be achieved when we take advantage of the environment information, which is the combined sewer system in this case. Thus, we optimize the advantage function through an optimization model based on the combined sewer system to further improve the control effect of all the RL agents.

During sampling process, we use some given rainfall data and the combined sewer system model to find the almost best action sequence by an optimization problem. We take the objective of reducing total CSO as an example (Eq. (4)).

$$\begin{aligned} &\min \sum_{t\in\{0,1,\dots,T\}} CSO_t \\ \text{s.t.}\quad &CSO_t, s_t = pipeline\_model(runoff_t, a_t, \theta) \\ &runoff_t = runoff\_model(rain_t, \mu) \end{aligned} \tag{4}$$

Where $CSO_t$ is the CSO volume in the time interval from $t$ to $t+1$, the $runoff\_model$ and $pipeline\_model$ are the models of combined sewer system with parameters $\theta$ and $\mu$. The $rain_t$ is the rain intensity in the time interval from $t$ to $t+1$, the time span has totally $T$ time interval. The $\{a_t\}, t \in \{0,1,\dots,T\}$ is the action sequence. The $\{s_t\}, t \in \{0,1,\dots,T\}$ is the corresponding system state.

The optimal action $\{\widehat{a_t}\}, t \in \{0,1,\dots,T\}$ and its corresponding state $\{\widehat{s_t}\}, t \in \{0,1,\dots,T\}$ can be obtained by solving this optimization model. Any solving algorithm can be applied. In this study, a basic genetic algorithm (GA) is used to solve Eq. (4). These optimal actions and states can be used to estimate a new maximum Q value (Eq. (5)), and then generate a new advantage function (Eq. (6)). Finally, we use this new advantage function to replace the original one in each RL models (including DQN, DDQN, PPO1, PPO2, A2C). This process of the advantage function improvement (AFI) is shown in Fig.4.

$$\max_{a_t} q(s_t, a_t, \theta) \approx q(\widehat{s_t}, \widehat{a_t}, \theta) = \mathbb{E}_t\left[\sum_{k=0}^{\infty}[\gamma^k r_{t+k+1}|\widehat{a_t}, \widehat{s_t}, \theta]\right] \tag{5}$$

$$\overline{A_t}(s_t, a_t, \theta) = q(s_t, a_t, \theta) - q(\widehat{s_t}, \widehat{a_t}, \theta) \tag{6}$$

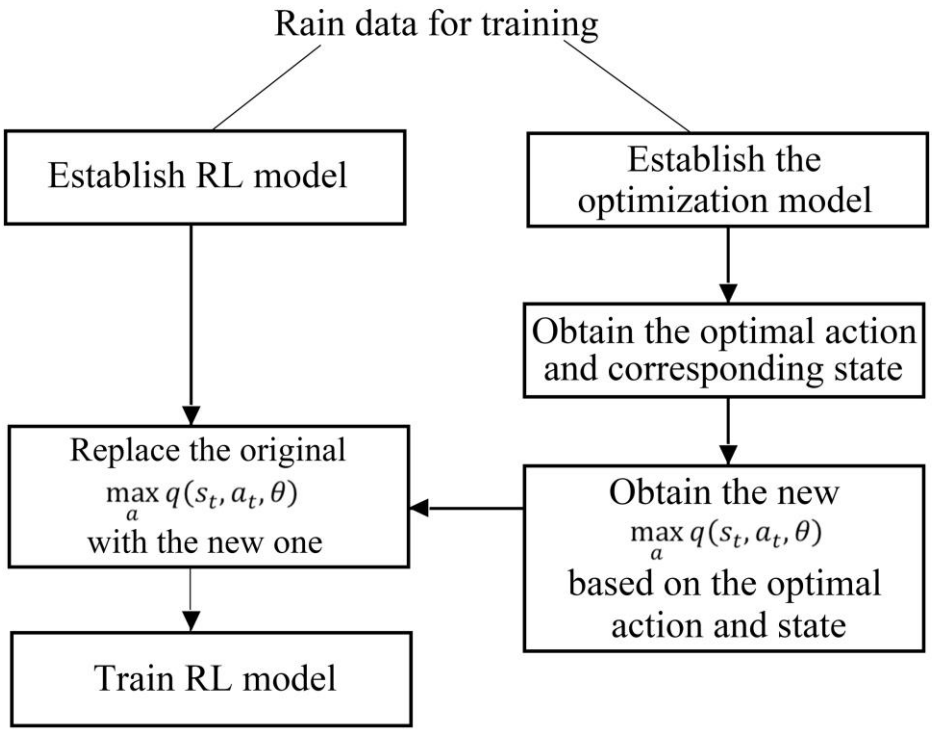



Fig.4 Advantage function improvement

These optimal actions and states give an estimation that is much closer to the maximum Q value of a given rainfall event, thus they are able to lead a better training effect. Also, the solving process of the optimization problem is decoupled from RL training, the only thing it provided is the estimation of the maximum Q value. Thus, the optimization problem can be computing on a parallel CPU.

We use the improvement rate (IR, Eq. (7)) as the indices of AFI performance. Where $CSO_{AFI,t}$ is the CSO volume of the RL models with AFI in the time interval from $t$ to $t+1$, $CSO_{non\_AFI,t}$ is the CSO volume of RL models without AFI in the time interval from $t$ to $t+1$. The $BL$ is the baseline of total COS volume, which can be provided through other RTC method or uncontrolled process. Thus, the $BL - \sum_{t\in\{0,1,\dots,T\}} CSO_{AFI,t}$ and $BL - \sum_{t\in\{0,1,\dots,T\}} CSO_{non\_AFI,t}$ mean the CSO reduction of AFI model and non AFI model compared to baseline. A large IR indicates a better performance of AFI.

$$IR = \frac{(BL - \sum_{t\in\{0,1,\dots,T\}} CSO_{AFI,t}) - (BL - \sum_{t\in\{0,1,\dots,T\}} CSO_{non\_AFI,t})}{(BL - \sum_{t\in\{0,1,\dots,T\}} CSO_{non\_AFI,t})} \quad (7)$$

### 3.3. Independent security system and the voting system

After training process, these five trained agents together with an MPC based security system are used to formulate our RTC system for safe control (Fig.5). In each

time step, all the five trained agents give their action reference to the same state of environment. Then, the security system predicts the rainfall of the next time-step, and test all of these five actions through an MPC model. After that, the action that satisfies the safety requirement and achieves the highest reward will be chosen as the control strategy of the next time-step. For easy understanding, this combined system, including the multi-RLs and the security system, is called as voting system in the rest of the paper, as it is similar to a voting process.

The safety requirements in the security system are problem-dependent. For instance, some drainage systems need a low water level in some of their nodes, or a reasonable load for pumps. Therefore, the requirements should be designed individually based on the system situations. In this case, the low water level of some nodes in the pipeline network is used as the safety requirements.

With this one-step checking, it is possible to choose a safe action, thus provide a guarantee of safety. Also, if all the actions given by these five agents are not safe enough, the system will provide a backup choice as the output action, such as water-level based action. Because the security system is decoupled from the RL framework, it offers an objective judgment on the given actions without influence from any of the agents, thus further ensure the reliability.

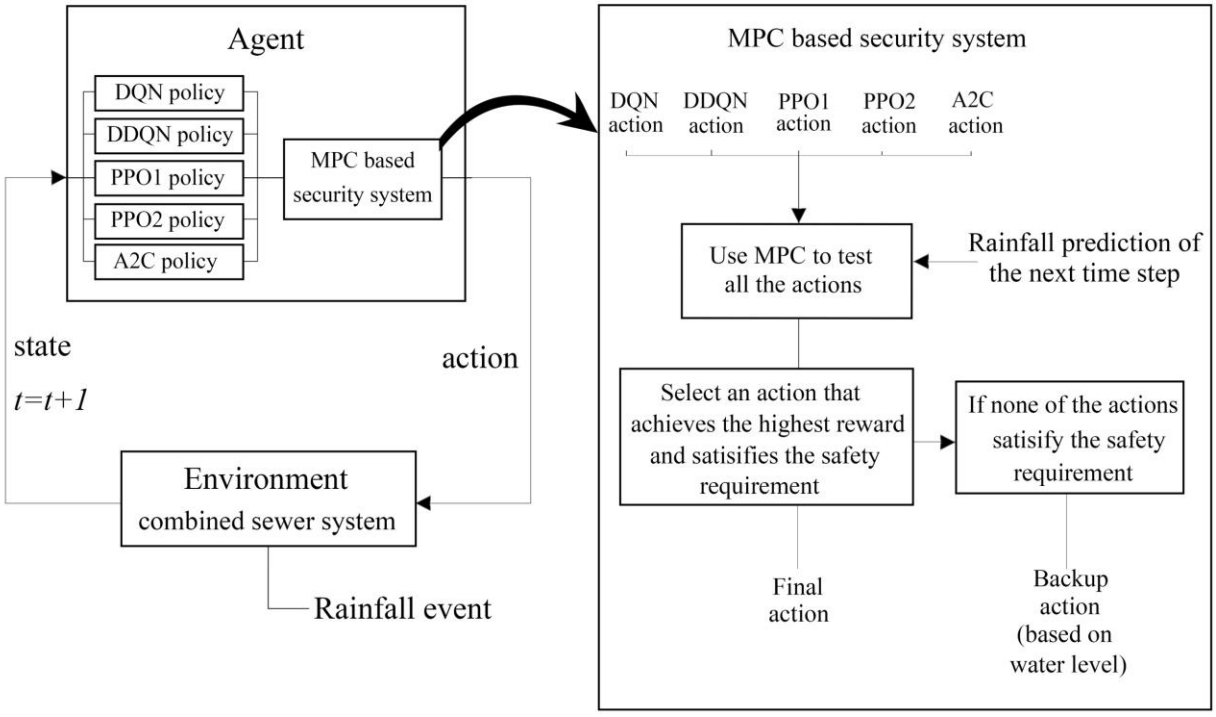



Fig.5 Security system and voting system

## 4. Case study

### 4.1. Combined sewer system of study area

The case study is the combined sewer system in a city in eastern China. It contains 211 nodes, 210 pipelines, and three pump stations, which includes C-pump station, K-pump station and R-pump station. C-pump station and K-pump station have one forebay and two pumps while the R-pump station has one forebay and four pumps. Considering the risk to property and public safety, a SWMM model of this combined sewer system is used as the environment. The schematic diagram of the model is shown in Fig.6. More details of this model can be found in our previous researches (Liao et al., 2019; Zhi et al., 2019).

According to Zhi et al. (2020), the areas that are vulnerable to flooding and overflow in this city (called high-risk areas) are located in the sub-catchments closed to the C and K pump station (red circles in Fig.6). Therefore, in this case, a reasonable safety requirement of the system operation can be defined as follow: The C, K-forebay, and the nodes in the high-risk sub-catchments (such as N1 and N2 in Fig.6) should keep a low water level during operation.

Currently, this combined sewer system has its own designed RTC system, which is water-level based. It sets a sequence of water-level threshold values, or set-points, to operate the pumps. The pump starts working if the water level of the forebay reaches its onset threshold and shuts down when the water level falls down to the shutoff threshold. The detailed onset/shutoff threshold values are given in Table 1. As the pumps drain water when the water level is high, this RTC system has the capability of reducing CSO at some level.

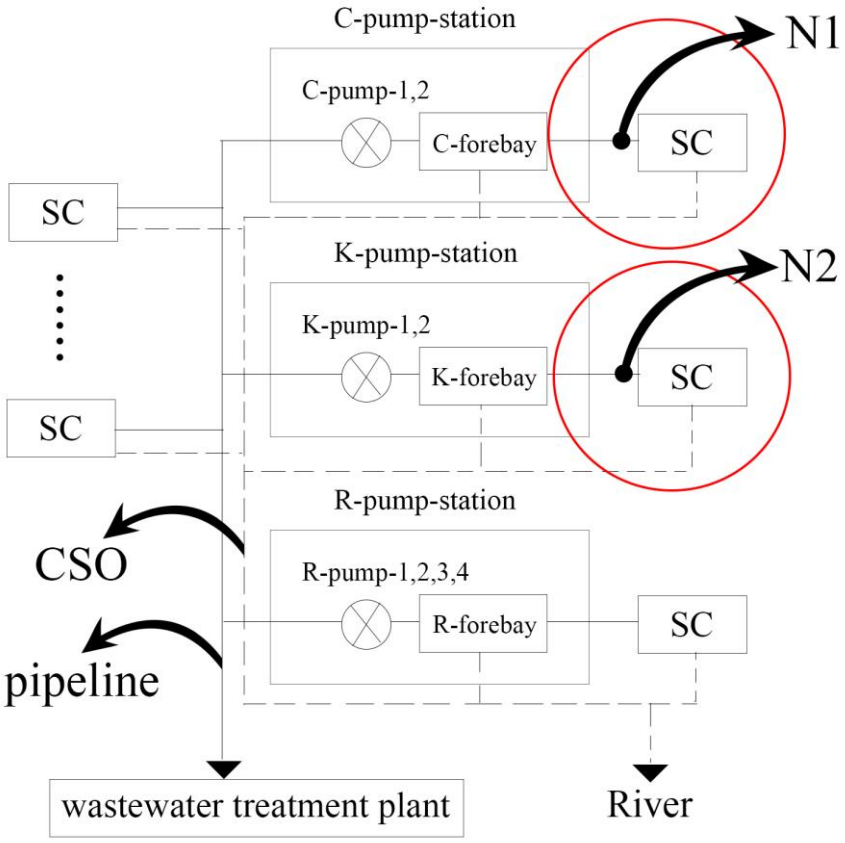



Fig.6. The schematic representation of the combined sewer system model. The SCs represent the sub-catchments. The high-risk areas are highlighted by red circles (Zhi et al., 2020). The N1 and N2 are two pipeline nodes in the high-risk area.

Table 1. The onset/shutoff threshold values of the water level based RTC.

| | Onset threshold (m) | Shutoff threshold (m) |
|---|---|---|
| C-pump-1 | 4.56 | 3.26 |
| C-pump-2 | 4.87 | 4.56 |
| K-pump-1 | 4.56 | 3.26 |
| K-pump-2 | 4.87 | 4.56 |
| R-pump-1 | 5.00 | 4.71 |
| R-pump-2 | 6.31 | 5.00 |
| R-pump-3 | 7.00 | 6.31 |
| R-pump-4 | 7.78 | 7.00 |

4.2. Rainfall events for training

As numerous rainfall events are required for RL systems training, a rain pattern formula (Eq. (8)) is employed to generate rainfall events.

$$q = \frac{A(1 + Clog(P))}{|tK - i|^n} \quad (8)$$

Where $q$ is the rainfall intensity, $i$ is a designed rainfall intensity, $A$ is the rainfall intensity with the recurrence period of one year, $C$ is an experience parameter, $P$ is the rainstorm return period, $t$ is the time, $K$ is the peak intensity position coefficient and the $n$ is a constant. To generate enough rainfall events for RL training, these parameters are randomly chosen within a range (Table 2) based on the historical research of the rainstorm intensity formula in the study area (Wang and Xu, 2016). A total of 1,200 rainfall events were generated and used for the agent training. Each rainfall event has a four-hour duration.

Table 2. The range of parameters in Eq. (8)

| $A$ (mm) | $C$ | $P$ (year) | $n$ | $i$ (mm/min) | $K$ |
|---|---|---|---|---|---|
| 21~35 | 0.939~1.20 | 1~5 | 0.86~0.96 | 16~22 | 0.3~0.8 |

## 5. Results

### 5.1. Configuration of the RL models

All the five RL models (DQN, DDQN, PPO1, PPO2, A2C) and the voting system are employed to establish RTC systems. As the control target is to reduce the total CSO during rainfall event, the state includes the current rainfall intensity, water level and water flow of forebays and volume of current total CSO. The action is an eight-dimensional vector which represents the control strategy of 8 pumps in the study area. The reward (Eq. (9)) is designed based on current CSO volume. The training process of each agent follows the steps in Section 3.1 with the rainfall data given in Section 4.2.

$$r_t = \begin{cases} -1 & if\ CSO\ volume\ in\ [t-1, t]\ is\ larger\ than\ 0 \\ 0 & else \end{cases} \quad (9)$$

After training, these five agents (named as DQN, DDQN, PPO1, PPO2, A2C) and the voting system are applied for the RTC of the case study during four designed rainfall events. The rainfall intensity of these four rainfall events are given in Fig.7. Moreover, the designed rainfall events are directly used as the rainfall prediction of the voting system in all the tests to eliminate the influence of inaccuracy of rainfall predictions. Considering the risk to property and public safety, all the tests are running on the combined sewer system model mentioned above. The control interval

is 10 min, which means the pumps are controlled every 10 min, thus all the computing processes in one time-step, including RL agents and the MPC of the security system, should be finished within this time limit.

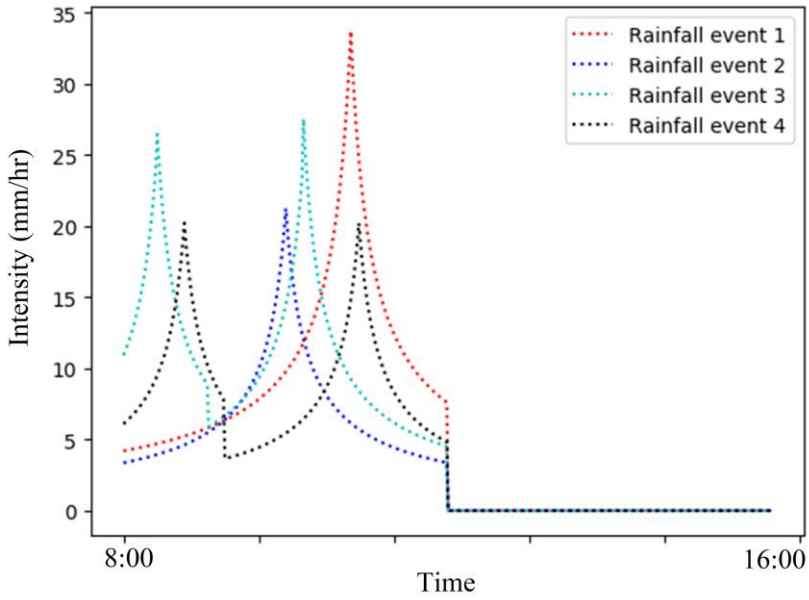



Fig.7 Four designed rainfall events used for the testing

According to the Section 4.1, the safety requirement of the voting system is defined as follow: the water level of C-forebay, K-forebay, N1, and N2 should lower than their safe line as much time as possible. The safe line of water level is set as 70% of the node depth. The depths and the safe line of each node are given in Table3. If the system is running under such a condition, we confirm that the controlling process is safe.

Table 3. The depth and the safe line of each node in the case study

| | C-forebay | K-forebay | N1 | N2 |
|---|---|---|---|---|
| Depth (m) | 5.63 | 5.8 | 2.21 | 1.957 |
| Safe line (m) | 3.941 | 4.06 | 1.547 | 1.3699 |

5.2. Efficiencies of the CSO reduction

The result about total overflow volume of these agents are given in Table 4. Their overflow volume at each time step during control process are given in Fig.8. For comparison, the CSO volume of the water-level based RTC system (given in Section

4.1) is also given. For easy understanding, we call it water level system in the rest of the paper. According to the results, all the RL models are able to show promise in CSO reduction compared to the water level system.

Table 4. Total overflow volume (10^3 $m^3$) of all the RL models during four rainfalls

| | DQN | DDQN | PPO1 | PPO2 | A2C | Voting | Water level system |
|---|---|---|---|---|---|---|---|
| Rain1 | 4.926 | 4.807 | 4.876 | 4.878 | 4.847 | 4.631 | 5.503 |
| Rain2 | 5.946 | 5.729 | 5.920 | 5.809 | 5.741 | 5.676 | 6.570 |
| Rain3 | 15.076 | 14.628 | 14.724 | 14.973 | 14.570 | 14.588 | 16.294 |
| Rain4 | 8.997 | 8.915 | 8.737 | 8.944 | 8.852 | 8.822 | 9.725 |

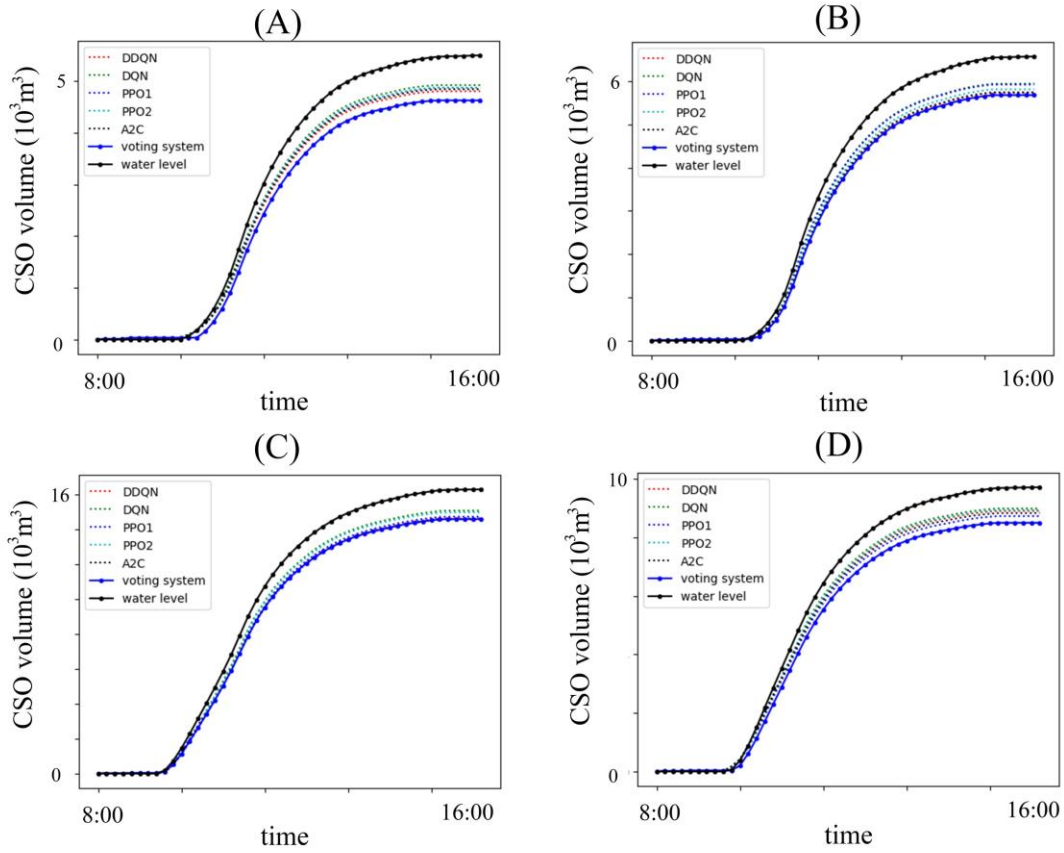



Fig.8 Total overflow volume of all the RL models at each time point during four rainfalls.

5.3. Advantage function improvement (AFI) on multi-RL models

We also employ the advantage function improvement to all the above methods for comparison. The corresponding results about total CSO volume and the improvement rate (IR, Eq. (7)) of them are given in Table 5. The baseline ($BL$ in Eq. (7)) are provided by the CSO volume of the water level system. Their overflow volume at

each time step during control process are given in Fig.9. The IR is in the range from 0.0% to 44.5%, which indicates that the AFI improves the CSO reduction for all the RL agents, except the DDQN in Rain2.

Table 5. Total CSO volume (10^3 $m^3$) and improvement rate (IR) of all the RL models with AFI in four rainfalls.

| | | Rain1 | Rain2 | Rain3 | Rain4 |
|---|---|---|---|---|---|
| DQN | CSO | 4.776 | 5.771 | 14.534 | 8.719 |
| | IR | 26.0% | 28.0% | 44.5% | 38.2% |
| DDQN | CSO | 4.776 | 5.729 | 14.531 | 8.717 |
| | IR | 4.5% | 0.0% | 5.8% | 24.4% |
| PPO1 | CSO | 4.805 | 5.82 | 14.707 | 8.73 |
| | IR | 11.3% | 15.4% | 1.2% | 0.7% |
| PPO2 | CSO | 4.778 | 5.741 | 14.881 | 8.839 |
| | IR | 16.0% | 8.9% | 7.0% | 13.4% |
| A2C | CSO | 4.806 | 5.727 | 14.547 | 8.505 |
| | IR | 6.3% | 1.7% | 1.3% | 39.7% |
| Voting | CSO | 4.629 | 5.642 | 14.505 | 8.508 |
| | IR | 0.2% | 3.8% | 4.9% | 34.8% |

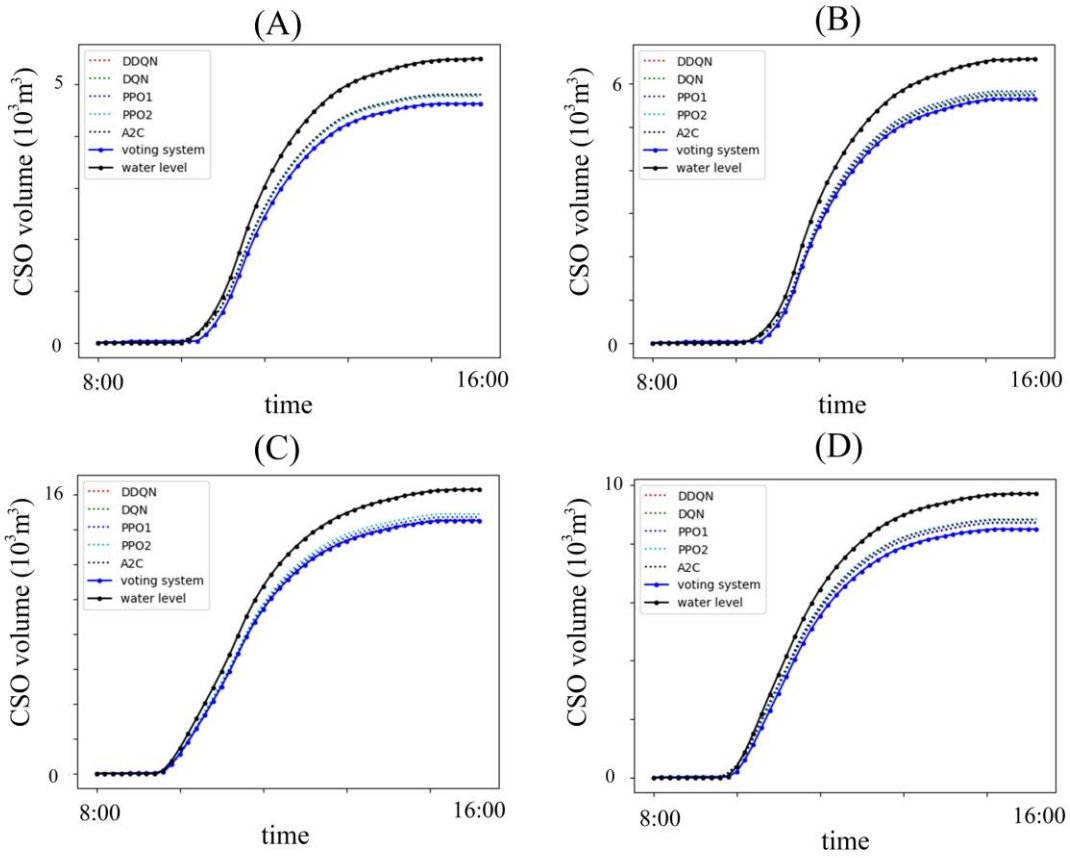
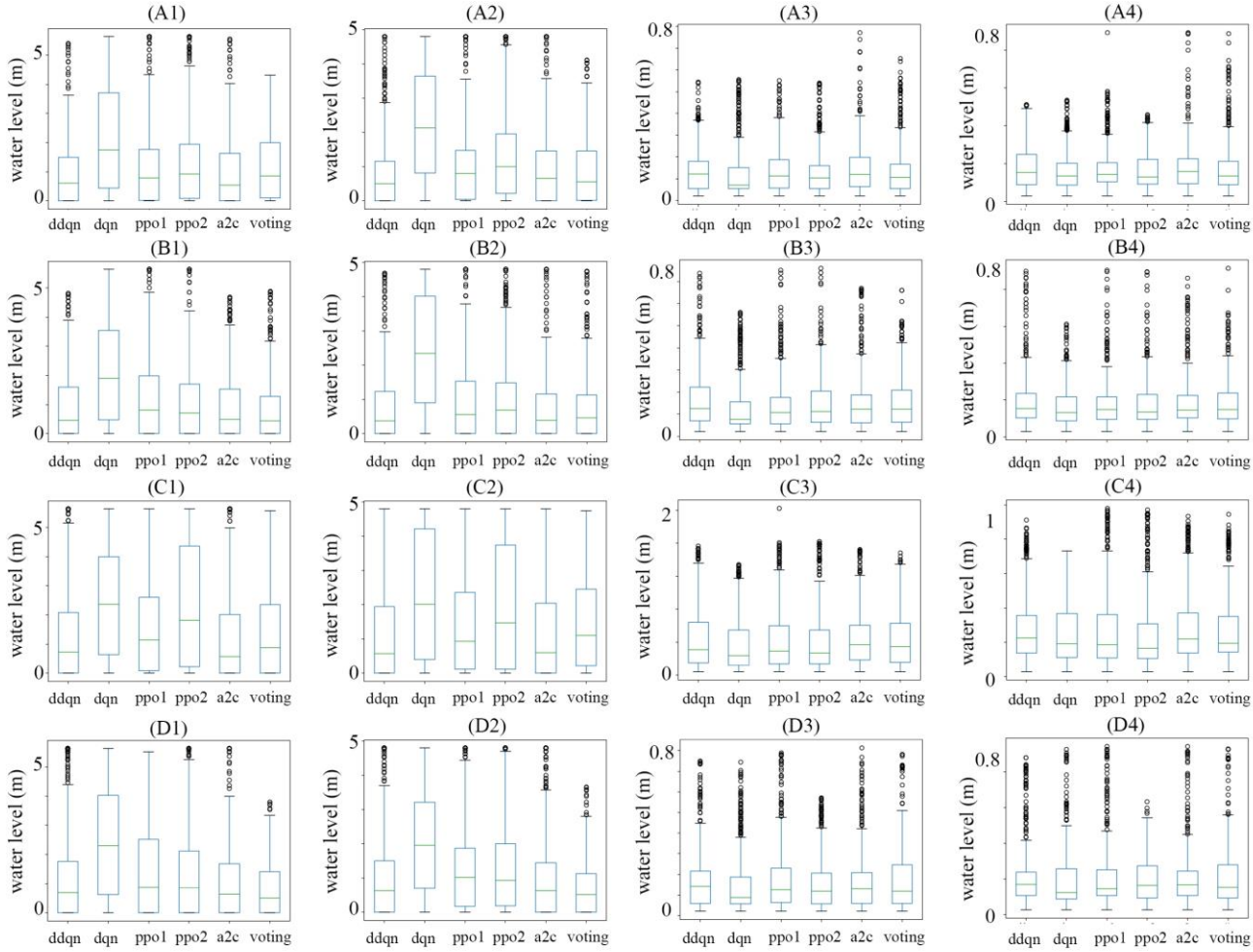



Fig.9 Total overflow volume of all the RL models (with AFI) at each time point during four rainfalls.

## 5.4. Safety of the voting system

To prove the safety of the voting system, the water level results of C forebay, K forebay, N1, and N2 during the rainfall duration of all the tests are given. If the water level of each node is lower than its safe line as much time as possible, then we confirm that the controlling process is safe. For each method, there are total of 480 water level data points for each node (C forebay, K forebay, N1, and N2) during an eight hours rainfall event (from 8:00 to 16:00). All of these data are presented through box plot in Fig.10 (the tests without AFI) and 11 (the tests with AFI).

Accordingly, the water level of DQN and PPO2 in the C forebay and K forebay surpass the safe line in Table 3 (3.941 m and 4.06 m), which means that the control process given by these two methods maybe unsafe for practically application. The water level of the voting system may not be the lowest all the time, it is more likely to stay at a low rank and satisfy the safety requirements in Table 3.




415

416 Fig.10 The water level of the C-forebay (1), K-forebay (2), N1 (3), and N2 (number 4)

417 during Rain1 (A), Rain2 (B), Rain3 (C), and Rain4 (D). Without AFI.

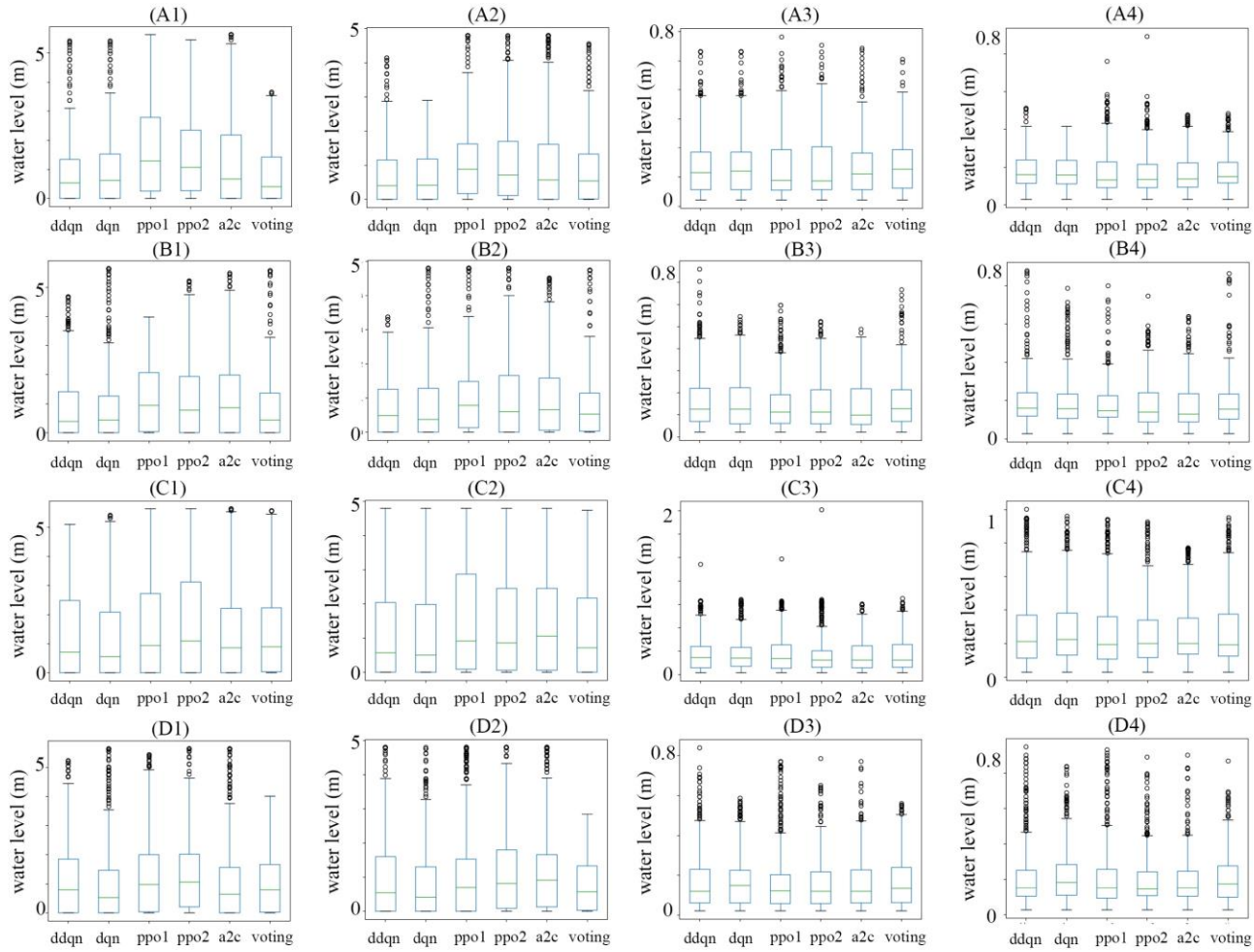



Fig.11 The water level of the C-forebay (1), K-forebay (2), N1 (3), and N2 (4) during Rain1 (A), Rain2 (B), Rain3 (C), and Rain4 (D). With AFI.

### 5.5. Computational cost

All the trainings and testing were run on a Windows Server (Intel ® Xeon® Gold 5117 CPU @2.00 GHz, RAM 32.0 GB). The training process of a single RL model took approximately 2-3 hours (with AFI) and 1-2 hour (without AFI). After training, each RL agent only needs around 0.01 s to generate action. The computing process of the security system at each time step took around 3 min, which is less than the 10 min used as control interval.

## 6. Discussion

### 6.1. CSO reduction and system optimization

#### 6.1.1. The AFI efficiency and the limitation of CSO reduction

From the above results, all the RL models reduce the overflow, which indicates that different types of RL models, policy-based or value-based, are effective in the

CSO control of the combined sewer systems. Also, the AFI achieves improvement of CSO reduction with the IR in the range from 0.0% to 44.5%, thus shows its efficient performance.

Meanwhile, according to Fig.8 and Fig.9, the difference among the RL agents in terms of CSO volume is getting smaller after the AFI technique is employed. The reason is that the AFI helps the RL agents reach the limits of CSO reduction. In fact, if the entire rainfall event can be accurately predicted and the optimization method is used to search for the optimal control strategy sequence, the obtained control strategy should be the global optimal solution and represents the limits of CSO reduction. From this perspective, the introduction of AFI is to provide information about the optimal solution may appear during the control process. Therefore, AFI helps all the RL systems improve their control effect, in other words, helps them to get closer to the limits of CSO reduction, which then leads to less difference among them.

#### 6.1.2. Local optimization

The "optimal solution may appear" here indicates the local optimal solution which is close to, rather than equal to, the limitation of CSO reduction. In practical RTC applications, we can neither accurately predict rainfall nor solve a complex optimization problem under the time constraints. Indeed, some researches sacrifice the accuracy to achieve a faster simulation (Xu et al., 2013; Lund et al., 2020). It means that if we want to plug the AFI into an online training process, it may lead to a local optimal control strategy, rather than the global optimal one. Therefore, the AFI can only be considered under off-line condition. How to ensure better optimality of AFI needs to be further explored and studied in the future.

#### 6.1.3. Optimization of voting system

According to the results, the voting system considers the optimal control strategy in each step of decision making, thus lead to a relatively better performance in the above examples. However, it is not necessarily optimal. For instance, in the Rain4 of the second test, the control effect of the voting system is relatively poor compared to A2C. The main reason is that the voting system can only guarantee the optimal of the selected action in each time step, rather than the optimization of the whole control process, as the optimal choice in each step may not definitely lead to the optimal of final results.

### 6.2. The safety of the voting system

Since the control process depends on the agent, which is a black-box model, its output is essentially probability-oriented and lacks the interpretability. Due to this, the safety of the RLC application is naturally questioned. For instance, based on the (C1) and (C2) of Fig.10, it may cause an unsafe situation if we only use DQN agent and PPO2 agent, as their water level is likely to stay at a high rank. In fact, both of these methods do not meet the safety requirement (given in Table 3) very well.

Compared to these RL models, the voting system is more likely to stay at a lower water level and satisfies the safety requirements in Table 3 as much time as possible. The main reason is that the voting system is able to avoid the risk caused by any one of the agents by selecting a safe action from all the given choices, therefore, it is safer than any single RL system during application. Although the causality between the control strategy and the system state cannot be explained in principle, the voting system provides a guarantee of safe operation to the control system.

### 6.3.Challenges

Although the computing time of each RL agent is very short (0.01 s), their training process is computationally expensive, especially after adding the AFI, as each step of training needs to solve an optimization problem synchronously. At present, the method to speed up the training process requires the use of more powerful computing equipment and parallel computing technology. According to the calculation process, the consumption of computing power mainly depends on two steps: environment simulation (SWMM model in this article) and RL training. How to optimize the calculation speed from these two aspects is what we are considering.

In addition, it needs to explain that the control strategy provided by RL agents only aims to minimize the overflow under the premise of a given rainfall event and a combined sewer system. If the entire sewer system is overloaded, simply relying on the RL control to adjust the hydrodynamics of the pipeline network cannot fundamentally reduce the overflow. This is one of the drawbacks faced by all the RTC methods.

## 7. Conclusion

Considering the safety and control effect improvement, a new RTC method based on multi-reinforcement learning, MPC, and an optimization model is introduced in this study. First, five individual agents are trained individually via five RL models. Then, an optimization model is applied to improve the advantage function of all the

RL agents. After that, an independent MPC based security system is established to ensure the safety of control strategy. Finally, our RTC method is established through the combination of these five agents and the independent security system.

The methodology and the case study show that: (i) Different RL methods are able to show promise in CSO reduction. (ii) The AFI technique imports the information of the optimal control strategy given by a corresponding optimization model, thus provide an improvement of CSO reduction with the maximum improvement rate of 44.5%. (iii) The security system is able to select a relatively better control strategy (or action) among all the actions given by agents and check its safety before use it. (iv) The voting system may lead to a relatively better control effect, but it is not necessarily optimal, because the optimal choice in each step may not definitely lead to the optimal of final results.

Meanwhile, our method is not a perfect solution. The AFI is suffered from the problem of local optimization. The training process of all the RL models are computationally expensive. Moreover, the control effect of RL control may be limited when the combined sewer system is overloaded, as it only provides a solution under the premise of a given rainfall event rather than extends the capacity of combined sewer system.

## Acknowledgements

This study was financially supported by the Major Science and Technology Program for Water Pollution Control and Treatment (grant no. 2018ZX07208006) and the National Natural Science Foundation of China (grant no. 51778451). We also thank the 111 Project (B13017) of Tongji University.

Data for this research, which is the calibrated SWMM model in the case study, is available in these references: Liao et al., (2019) and Zhi et al., (2019).

## Reference

Abhiram Mullapudi, Matthew J. Lewis, Cyndee L. Gruden, Branko Kerkez. (2020). Deep reinforcement learning for the real time control of stormwater systems, Advances in Water Resources, Volume 140.

Castelletti A, Pianosi F, Restelli M. (2013). A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier

approximation in a single run. Water Resources Research, 49(6): 3,476–3,486.

Castro, D. D., Tamar, A., & Mannor, S. (2012). Policy Gradients with Variance Related Risk Criteria. international conference on machine learning.

Congcong S., Luis R., Bernat J., Jordi M., Eduard M., Ramon G., Montse M., Vicenç P., Gabriela C. (2020). Integrated pollution-based real-time control of sanitation systems. Journal of Environmental Management. Volume 269, 2020, 110798, ISSN 0301-4797

Fu G, Butler D, Khu S T. (2008). Multiple objective optimal control of integrated urban wastewater systems. Environmental Modelling & Software, 2008, 23(2): 225-234.

Garcia, J., & Fernandez, F. (2015). A comprehensive survey on safe reinforcement. learning. Journal of Machine Learning Research, 16(1), 1437-1480.

Geibel, P., & Wysotzki, F. (2005). Risk-sensitive reinforcement learning applied to control under constraints. Journal of Artificial Intelligence Research, 24(1), 81-108.

Gehring, C., & Precup, D. (2013). Smart exploration in reinforcement learning using absolute temporal difference errors. adaptive agents and multi agents systems.

Gu X, Liao Z, Zhang G, et al. (2017). Modelling the effects of water diversion and combined sewer. overflow on urban inland river quality. Environmental Science and Pollution Research, 2017, 24(26): 21038-21049.

Joseph-Duran B, Ocampo-Martinez C, Cembrano G. (2015). Output-feedback control of combined sewer networks through receding horizon control with moving horizon estimation. Water Resources Research, 51(10): 8,129-8,145.

Kerkez, B., Gruden, C., Lewis, M.J., Montestruque, L., Quigley, M., Wong, B.P., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A., Pak, C. (2016). Smarter Stormwater Systems. Environ. Sci. Technol. 50 (14).

Labadie J W. (2014). Advances in Water Resources Systems Engineering: Applications of Machine Learning. Modern Water Resources Engineering. Humana Press, Totowa, NJ: 467–523.

564 Liao, Z., Gu, X., Xie, J. et al. (2019). An integrated assessment of drainage system
565 reconstruction based on a drainage network model. Environ Sci Pollut
566 Res 26: 26563.
567 Lund N S V, Falk A K V, Borup M, et al. (2018). Model predictive control of urban
568 drainage systems: A review and perspective towards smart real-time water
569 management. Critical reviews in environmental science and technology,
570 2018, 48(3): 279-339.
571 Lund, N. S. V., Borup, M., Madsen, H., Mark, O., & Mikkelsen, P. S. (2020). CSO
572 reduction by integrated model predictive control of stormwater inflows: A
573 simulated proof of concept using linear surrogate models. Water Resources
574 Research, 56, e2019WR026272.
575 Madani K, Hooshyar M. (2014). A game theory–reinforcement learning (GT–RL)
576 method to develop optimal operation policies for multi-operator reservoir
577 systems. Journal of hydrology, 519: 732–742.
578 Mailhot A., Talbot G., Lavallée, B. (2015). Relationships between rainfall and
579 Combined Sewer Overflow (CSO) occurrences. Journal of Hydrology,
580 2015, 523:602-609.
581 Minh, V., Kavukcuoglu, K., Silver, D., Rusu, A.a., Veness, J., Bellemare, M.G.,
582 Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S.,
583 Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra,
584 D., Legg, S., Hassabis, D. (2015). Human-level control through deep
585 reinforcement learning. Nature 518, 529–533.
586 Minh, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T., Silver, D.,
587 Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement
588 learning. international conference on machine learning.
589 Moldovan, T., & Abbeel, P. (2012). Safe Exploration in Markov Decision
590 Processes. arXiv: Learning.
591 Ochoa, D., Rianobriceno, G., Quijano, N., & Ocampomartinez, C. (2019). Control of
592 Urban Drainage Systems: Optimal Flow Control and Deep Learning in
593 Action. advances in computing and communications.
594 Pablo Quintia Vidal, Roberto Iglesias Rodriguez, Miguel Rodriguez Gonzalez, and
595 Carlos V azquez Regueiro. (2013). Learning on real robots from

experience and simple user feedback. Journal of Physical Agents, 7(1), ISSN 1888-0258.

Pan, X., You, Y., Wang, Z., & Lu, C. (2017). Virtual to real reinforcement learning for autonomous driving. arXiv preprint arXiv:1704.03952.

Rauch W, Harremoes P. (1999). Genetic algorithms in real time control applied to minimize transient pollution from urban wastewater systems. Water research, 1999, 33(5): 1265-1277.

Schütze M, Campisano A, Colas H, et al. (2002). Real-time control of urban wastewater systems-where do we stand today? Global Solutions for Urban Drainage. 2002: 1-17.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv: Learning.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., & Moritz, P. (2015). Trust Region Policy Optimization. international conference on machine learning.

Sebastian Peitz, Stefan Klus. (2019). Koopman operator-based model reduction for switched-system control of PDEs. Automatica. 106, 184-191.

Shao K, Zhu Y, Zhao D. (2018). StarCraft Micromanagement with Reinforcement Learning and Curriculum Transfer Learning. IEEE TransActions on Emerging Topics in Computational Intelligence.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., & Chen, Y. (2017). Mastering the game of go without human knowledge. Nature, 550(7676), 354.

Suarez J, Puertas J. (2005). Determination of COD, BOD, and suspended solids loads during combined sewer overflow (CSO) events in some combined catchments in Spain. Ecological Engineering, 24(3): 199–217.

Sutton R S, Barto A G. (2018). Reinforcement learning: An introduction[M]. MIT press.

Sutton, R. S., Mcallester, D., Singh, S., & Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. neural information processing systems.

Tan Zhiyong, Chai Quek, Philip Y.K. Cheng. (2001). Stock trading with cycles: A financial application of ANFIS and reinforcement learning. Expert Systems with Applications, 38(5):4741-4755.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-Learning. national conference on artificial intelligence.

Van Hasselt. (2010). Double Q-learning. Advances in Neural Information Processing Systems, 23:2613–2621.

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., & De Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. international conference on machine learning.

Wan P, Lemmon M D. (2007). Distributed flow control using embedded sensor-actuator networks for the reduction of combined sewer overflow (CSO) events. 2007 46th IEEE Conference on Decision and Control. IEEE: 1,529–1,534.

Wang Rui, Xu Deqian. (2016). Derivation of Rainstorm Intensity Formula for Hefei City. Journal of China Hydrology, 2016, 36(1): 71-74.

Wu, B. (2019). Hierarchical macro strategy model for moba game ai. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 1206-1213.

Xie J, Chen H, Liao Z, et al. (2017). An integrated assessment of urban flooding mitigation strategies for robust decision making. Environmental modelling & software, 2017, 95: 143-155.

Xu Z, Liao Z. (2013). A Systematic View Is Key: The Successful Case of Suzhou Creek Rehabilitation. Environmental Science & Technology, 47(21): 11,936–11,937.

Xu, M., Van Overloop, P. J., & De Giesen, N. C. (2013). Model reduction in model predictive control of combined water quantity and quality in open channels. Environmental Modelling and Software, 72-87.

Yong Song, Yi bin Li, Cai hong Li, and Gui fang Zhang. (2012). An efficient initialization ap- proach of q-learning for mobile robots. International Journal of Control, Automation and Systems, 10(1):166–172.

Zhi Guozheng, Liao Zhenliang, Wenchong Tian, Jiang Wu. (2020). Urban flood risk assessment and analysis with a 3D visualization method coupling the PP-PSO algorithm and building data. Journal of Environmental Management. Volume 268, 110521, ISSN 0301-4797.

661 Zhi G, Liao Z, Tian W, Jiang W. (2019). A 3D dynamic visualization method coupled
662 with an urban drainage model. Journal of. Hydrology, 577: 123988.